Task Group on Series Numbering

Standing Committee on Automation
Program for Cooperative Cataloging

Final report

Sherman Clarke, New York University
Greta de Groat, Stanford University
Stephen Hearn, University of Minnesota
Gary L. Strawn, Northwestern University, chair

3 August, 2002

Table of contents

Appendixes

Letter of transmittal

The Task Group on Series Numbering created by the Standing Committee on Automation of the Program for Cooperative Cataloging was asked to examine the conditions that prevent series headings from being arranged by automated systems in numerical order, and to identify an algorithmic approach for the better arrangement of series headings. Having brought its investigations to a close, the task group is pleased to submit the accompanying report, which contains a description of its working methods, experiments and findings.

In the course of its work, the task group identified a number of characteristics of series numbering that prevent the perfect sorting of series headings. The task group was not able to find an algorithmic solution for the problems caused by these characteristics; the task group describes these characteristics here, so that appropriate parties can be asked to consider the possibility of changes. These characteristics of series numbering have been an accepted part of cataloging practice for decades, and there is little hope that existing bibliographic records could ever be modified to improve the order in displays of the series headings they contain. However, the task group believes that changes to the manner in which series numbering is recorded can be considered for newly created bibliographic records, so that the series headings in those new records will have a reasonable chance of being arranged correctly. Since headings bearing these characteristics cannot now be sorted into numerical order, few new disruptions to the order of series headings should occur if practices are changed in mid-course. The task group also includes here suggestions for related changes to practice for use of the authority 642 field and the manner in which library systems employ the 642 field. The task group believes that the benefits of these changes are self-evident, and urges their adoption. If these changes are not made, the sorting of series headings will continue to be a problem that can be solved by any library system with at best partial satisfaction.

- Series numbering with roman numerals cannot be sorted algorithmically. Descriptive cataloging practices should be redesigned to indicate that if a roman numeral appears in the series numbering, the series statement must appear in a 490 field,[1] and the series access point—with the roman numeral replaced by an arabic numeral—must appear in an 8XX field.[2]

  Instead of this construction:
  ```
     440  0 ǂa Series heading ; ǂv III, 2
  ```
  Use this construction:
  ```
     490 1  ǂa Series heading ; ǂv III, 2
     830  0 ǂa Series heading ; ǂv 3, 2.
  ```

[1] Or in the equivalent of a 490 field. For example, the series statement containing the roman numeral for a microform might be carried in subfield ǂf of a 533 field.

[2] A note to section 1.1 of the report contains additional comments on roman numerals and related descriptive cataloging conventions.

Using a pair of fields tagged 490 and 830 as suggested here preserves for full-record displays the hierarchy implied by the different types of numerals, while also providing a form of the heading that can be sorted correctly. An alternative (somewhat less clear) would be further to qualify the notion that the 440 field must be a literal transcription of information found in the item being cataloged, and to use arabic numerals in the 440 field when roman numerals appear in the item.

- Numbers (including ordinal numbers) intended to be arranged numerically should be represented by digits, not words, and also not by combinations of digits and alphabetic characters. Descriptive cataloging practices should be redesigned to indicate that if a 'number' in series subfield $v contains a combination of digits and non-digits, or is represented without digits at all, the series statement must appear in a 490 field and the series access point, with the numbers represented solely by digits, must appear in an 8XX field.

Instead of this construction:

```
440  0 ǂa Werken uitgegeven door de Faculteit van
        de Letteren en Wijsbegeerte ; ǂv 167e afl.
```

Use this construction:

```
490 1  ǂa Werken uitgegeven door de Faculteit van
        de Letteren en Wijsbegeerte ; ǂv 167e afl.
830  0 ǂa Werken uitgegeven door de Faculteit van
        de Letteren en Wijsbegeerte ; ǂv 167. afl.[3]
```

- Years represented by two digits should be extended to four digits. Descriptive cataloging practices should be redesigned to indicate that if a year in subfield ǂv is represented by two digits, the series statement must appear in a 490 field and the series access point, with the date expanded to include the century digits, must appear in an 8XX field.[4]

Instead of this construction:

```
440  0 ǂa PRIO report ; ǂv 87/3
```

Use this construction:[5]

```
490 1  ǂa PRIO report ; ǂv 87/3
830  0 ǂa PRIO report ; ǂv 1987/3.
```

- Series numbers that include thousands separators[6] should be recorded in subfield ǂv without the separators.

[3] Or 'afl. 167' or even just '167'.

[4] Cf. LCRI 21.30L, 'Numbering consisting of an indication of a year and sequential number within a year.'

[5] In this example, the cataloger has determined that '87' represents the year 1987. A note to section 2.7 of the report describes some of the difficulties inherent in the automatic conversion of two digits into a four-digit year.

[6] In the United States, the comma is used as the thousands separator. Other conventions are used elsewhere.

Instead of this construction:[7]

```
830  0  ǂa 20th-century legal treatises ; ǂv
        fiche 4,293-4,296.
```

Use this construction:

```
830  0  ǂa 20th-century legal treatises ; ǂv
        fiche 4293-4296.
```

- If the members of a multi-part item are not numbered consecutively within the series, the bibliographic record should contain a separate series access point for each consecutive group of numberings.[8]

Instead of this construction:

```
490 1  ǂa 10/18 ; ǂv 971-972, 991-992, 1008
830  0 ǂa 10/18 ; ǂv 971-972, etc.
```

Use this construction:

```
490 1  ǂa 10/18 ; ǂv 971-972, 991-992, 1008
830  0 ǂa 10/18 ; ǂv 971-972.
830  0 ǂa 10/18 ; ǂv 991-992.
830  0 ǂa 10/18 ; ǂv 1008.
```

- If subfield ǂv contains more than one level of hierarchy in the numbering, the elements should be given in order of precedence, from broadest to narrowest.

- Dates consisting of year plus month and day, year plus season, etc., should be recorded with the year before the month or season and the month before the day.

Instead of this construction:

```
440  0 ǂa Proceedings of the Eisenhower Medical
        Center ; ǂv winter 1980
```

Use this construction:

```
490 1  ǂa Proceedings of the Eisenhower Medical
        Center ; ǂv winter 1980
830  0 ǂa Proceedings of the Eisenhower Medical
        Center ; ǂv 1980 winter.
```

- Designations such as ‘new series’ and ‘3rd series’ should always be carried in subfield ǂn or ǂp and not (as is done at present) sometimes in subfield ǂn or ǂp, sometimes in subfield ǂv.[9]

[7] This example presumes the existence of this 533 field: `ǂa Microfiche. ǂb Woodbridge, Conn. : ǂc Primary Source Media, ǂd 1995. ǂe 4 microfiches. ǂf (20th-century legal treatises ; fiche 4,293-4,296)`

[8] Cf. AACR2R 1.6G2.

[9] Cf. AACR2R 1.6H3. Unless, of course, the ‘series’ number *is* the numbering of the series.

Instead of this construction:

```
440  0 ǂa Marian Library studies ; ǂv new ser.,
       v. 12
```

Use this construction:

```
440  0 ǂa Marian Library studies. ǂp New series ;
       ǂv v. 12
```

Instead of this construction:

```
440  0 ǂa eiträge zur Wissenschaft vom alten und
       neuen Testament ; ǂv 3. Folge, Heft 2
```

Use this construction:

```
440  0 ǂa Beiträge zur Wissenschaft vom alten und
       neuen Testament. ǂp 3. Folge ; ǂv Heft 2
```

- Library systems should be redesigned to apply information in the 008 and 642 fields of authority records when verifying bibliographic series headings.[10] The operator should be warned by the system if the form of the data in subfield ǂv does not correspond to information in the authority record. (Because the automated evaluation of subfield ǂv cannot be performed with absolute reliability, the library system should not prevent the operator from storing a bibliographic record whose series numbering does not correspond to information in a series authority record.)

  Series numbering: `ǂv 19`
  Series numbering example from the 642 field of a series authority record:
  `ǂa Heft 24`
  The series numbering example indicates that ǂv should contain the text 'Heft' plus a number, but the ǂv contains only a number. The library system should warn the operator that the bibliographic series numbering does not correspond to the numbering example.

  Interesting problems surround the use of the authority 642 field to detect problems with series numbering. Although much of value can be extracted the existing 642 field, the field would be even more useful in automated validation were patterns of its use to change, or if the field were redesigned altogether.

One problem—which exists at least in the theoretical realm[11]—remains unsolved. The problem would arise when these conditions are met:

- A series is numbered

[10] Appendix E to the report describes a simple scheme for this test. The appendix also contains further recommendations regarding the 642 field. The verification of headings is a process separate from the process of normalization of series numbering described in the body of the report.
[11] No example of this problem that did not unambiguously represent improper cataloging could be found among the 696,510 series headings used by the task group for testing.

- The same series has numbered parts (which may or may not have their own numbering)
- The numbering for the part normalizes to the same form as the numbering for the main series (e.g., both normalize to numerals)

Under these conditions, entries for the numbered subseries could (under some of the normalization schemes proposed in this report) fall between entries for the main series in a list of the members of the series:

```
Heading ; ǂv no. 1
Heading. ǂn 2, ǂp Bibliography ; ǂv no. 3
Heading ; ǂv no. 5
```

Such a problem could be eliminated (if it is deemed worthy of solution) were some appropriate gimmick employed in the normalized form of the heading. This gimmick would cause all of the members of the basic series to be arranged before any members of the series with numbered parts.

```
Heading ; ǂv no. 1
Heading ; ǂv no. 5
Heading. ǂn 2, ǂp Bibliography ; ǂv no. 3
```

Finally, the task group notes that some of the techniques described in its report for the normalization of subfield ǂv of series headings could with profit be extended to other subfields that contain information that might be expected by catalog users to be arranged in numerical order. This would allow additional fields to be arranged correctly in displays, without requiring changes to existing records. Similarly, the task force feels that its recommendations concerning the recording of information—such as the elimination of roman numerals—could with profit be extended to such other subfields. (Perhaps it is time for a consideration of the presentation of numbers in all access fields.[12]) Subfields that would benefit from more sophisticated normalization include:

- Subfield ǂn in conference headings, in uniform title headings, and in the title portion of name/title headings[13]

```
Alabama Symposium on English and American Literature
   ǂn (5th : ǂd 1978 : ǂc University of Alabama)
```

[12] It would be interesting to know the effect that would be produced on the order of headings if the technique described in Appendix C of the report were applied to digits in all access fields, including fields not directly subject to authority control, such as the 245 field. Converting all digits to a standard form, rather than only some of them, would probably simplify the corresponding changes required in search algorithms.

[13] As of May 6, 2002, the database of Northwestern University Library held 142,803 bibliographic records (out of about 3.2 million) with subfield ǂn in conference headings and titles; these records contained a total of 167,036 headings with subfield ǂn. (An indication of the prevalence of this subfield in this database is included here simply to allow readers to gauge the scope of the problem such subfields present.)

```
Alabama Symposium on English and American Literature
   ǂn (9th : ǂd 1982 : ǂc University of Alabama)
Alabama Symposium on English and American Literature
   ǂn (10th : ǂd 1983 : ǂc University of Alabama)

Mahler, Gustav, ǂd 1860-1911. ǂt Symphonies, ǂn no. 1,
   ǂr D major
Mahler, Gustav, ǂd 1860-1911. ǂt Symphonies, ǂn no. 4,
   ǂr G major
Mahler, Gustav, ǂd 1860-1911. ǂt Symphonies, ǂn no. 10
```

- Subfield ǂp in ‘Bible’ headings that contain chapter and verse designations.[14] If the task group’s suggestions for the treatment of roman numerals is not followed and roman numerals are retained in headings, the development of program code to identify and convert roman numerals appearing in this subset of headings into arabic numerals for sorting would be well repaid.[15]

```
Bible. ǂp O.T. ǂp Genesis XI, 26-XX, 18
Bible. ǂp O.T. ǂp Genesis XII-L. ǂl English. ǂs New
   English. ǂf 1978.
Bible. ǂp O.T. ǂp Genesis XVIII, 1-XXII, 24
Bible. ǂp O.T. ǂp Genesis XXVI-L.
Bible. ǂp O.T. ǂp Genesis XLI, 1-XLIV, 17
```

[14] As of May 7, 2002, the bibliographic records in Northwestern University Library’s database held 4,985 ‘Bible’ headings with a subfield ǂp containing numbering for chapter (and verse). The task force notes that moving chapter and verse information for ‘Bible’ headings from subfield ǂp to subfield ǂn, where similar information is carried for all other headings, would also be an improvement.

[15] The conversion of roman numerals in series subfield ǂv into digits without inadvertently converting other text composed of the same characters is an impossible task; but, because of the restricted context, it should be possible reliably to convert roman numerals in subfield ǂp of ‘Bible’ headings into digits for sorting. Note that numberings in ‘Bible’ headings that refer to a range of chapters should be normalized in some way that prevents them from falling between headings that refer to chapter and verse; in these headings, the punctuation can probably not be replaced in the normalized form by a single space.

```
   Bible. ǂp N.T. ǂp John I, 13
   Bible. ǂp N.T. ǂp John I-XII
   Bible. ǂp N.T. ǂp John I-XV
```

Summary

Headings for the members of a numbered series should be presented in library catalogs in numerical order. Unfortunately, substantial difficulties impede the realization of this goal. The current versions of most automated library systems make no attempt to overcome these difficulties, and do not provide useful displays of numbered series headings. A task group appointed by the Standing Committee on Automation of the Program for Cooperative cataloging has identified a number of algorithms that could improve the sequencing of series headings by creating an improved normalized form of the series numbering. These algorithms vary in the amount of programming effort required, the degree to which system performance would be affected by their use, and the degree to which they achieve the proper sorting of series headings.

There appears to be a direct relationship between the amount of detail built into the algorithm for normalizing series numbering and the correctness of the result. The more effort spent designing and programming routines to sort series headings, the better will be the order of the sorted headings. None of the algorithms proposed in the task group's report can produce perfect sorts in every case, because the data with which the algorithms must work are not perfect. The most elaborate of the algorithms can come quite close to the ideal, but even the simplest provides for markedly better arrangement of series headings than that provided by most of today's automated library systems.

The approaches identified by the task group for normalizing series numbering are:

- Expand numerals to a fixed length and perform standard normalization on the remainder of the numbering
- Expand numerals to a fixed length, remove characters that appear before the first digit, and perform standard normalization on the remainder of the numbering
- Expand numerals to a fixed length, remove those characters appearing before the first digit that seem to constitute caption information, and perform standard normalization on the remainder of the numbering
- Expand numerals to a fixed length, remove characters that seem to constitute caption information, regularize some information, and perform standard normalization on the remainder of the numbering

Additional approaches may be devised. The course settled on by library system vendors and their customers will vary, depending on the perceived seriousness of the problem and the programming resources available. All parties are urged to implement one of the suggested techniques, or some similar technique, to improve the order of series headings in catalog displays.

1 Introduction

1.1 Background

Records in library catalogs were once presented in card form, and the cards arranged by filers. Filers performed their work by comparing information on a new card to information on cards already in place until the right location for the new card was determined. Filers were able to find the right location for each card in part because a set of rules guided their interpretation of the information in front of them, and in part because experience led them to overlook small inconsistencies among the data on cards.

The shift from cards to machine-readable records and from human filers to computer processing changed the nature of the task of arranging records, and the manner in which it is performed. The library system extracts access points from records, normalizes them in some manner to remove the effects of variations in capitalization and punctuation, and (typically) stores the normalized headings in a table.[1] Once the normalization routine has been written the question of arrangement has been settled; normalized entries appear in the order dictated by a computer's collating sequence. Filing rules have in this manner been supplanted by rules for the normalization of headings; normalized headings are simply sorted character by character. The design of the normalization routine therefore becomes the key factor affecting the arrangement of headings in automated library catalogs.

Numbered series[2] constitute one class of heading whose arrangement in automated library systems is less satisfactory than was common in pre-computer catalogs.[3] That numbered series headings should be presented in numerical order[4] has been an axiom of library filing rules for many decades, and card catalogs arranged by filers could claim to achieve this ideal. However, the variety of ways in which series numbering may be presented in bibliographic records confronts the designer of a normalization routine with substantial difficulties—difficulties that might seem on their face to preclude the

[1] Although the typical automated library system uses the normalized forms of headings for retrieval and the arrangement of displays, the normalized headings themselves are often not visible to users of the system. See section 2.6 for a discussion of the competing demands placed on the normalized heading by retrieval and sorting operations.

[2] Numbered series are series whose access points contain subfield ǂv. The term *numbering* is used for the contents of subfield ǂv even if does not contain any digits.

[3] An informal survey of six major vendors of library automation systems during the ALA conference at Atlanta in June, 2002 revealed one system that applies what is here called level 1 normalization to series subfield ǂv, one that applies system normalization, one that sorts subfield ǂv in numeric order if the field contains only digits and applies system normalization in other cases, and three systems that make no attempt to arrange series by number. A seventh system, investigated later, also does not attempt to sort series by number but provides a 'follow-on' search that can in theory be used to restrict results to a single member of a series; but the manner in which this second search is performed produces many false matches.

[4] Card catalogs presented the entries for a series in *ascending* numerical order. The assumption implicit throughout this report is that automated library systems should likewise present series headings in ascending order. If it were desired instead to present the headings in *descending* order (placing, at least much of the time, the most recent volumes first), suitable adjustments would have to be made to the normalization schemes described in this report.

possibility of arranging series entries in numerical order. The following paragraphs describe some of these difficulties.

- Captions that accompany the numbers are not presented by publishers, and are not recorded by catalogers, with rigorous consistency.[5] If captions are used as found in subfield ǂv to arrange entries, any inconsistency in the caption will lead to entries being presented in effectively random numerical order.

```
ǂv 5[6]
ǂv no. 7
ǂv no. 8
ǂv v. 1
ǂv v. 2-3
ǂv v. 9
ǂv v13
ǂv vol. 4

ǂv 10. Abt., 3. T., 3. Bd.
ǂv 2A, 4T
```

- The grammar of some languages allows the caption to appear either before or after the number. (When the number appears before the caption, the number is often, either explicitly or implicitly, an ordinal number.) Headings in different bibliographic records for members of a single series often follow different practices.

```
ǂv 1. Bd.
ǂv 9. Bd.
ǂv Bd. 5
ǂv Bd. 8

ǂv 5th v.
ǂv v. 2
```

- The numbering may be presented in roman instead of arabic numerals[7].

[5] The caption in subfield ǂv of a series heading is supposed to be controlled by the series numbering example in the 642 field of the authority record for the heading. Unfortunately, most automated library systems do not make use of information in the 642 field to notify operators of inconsistencies in the captions of headings in bibliographic records. For a fuller discussion of this situation, see Appendix E.

[6] Here and elsewhere, the entries in each group represent series numberings used with bibliographic instances of a single series heading.

[7] AACR2R Appendix C.2B1 instructs the cataloger to replace roman numerals in series numbering with arabic numerals, unless (C.2B2) the substitution makes the series statement 'less clear.' Essentially the same instruction appears in Appendix IV.A of both the first edition of AACR (1967) and *Rules for descriptive cataloging* (1949). Although *Catalog rules, author and title entries* (1908) does not explicitly cover this point, the examples included with rule 166 (p. 54) suggest that it was common practice at one time to allow roman numerals in series numbering when the item being cataloged used them, even if no ambiguity would result from their conversion to arabic. It is in any case true that series numberings in many

```
ǂv L
ǂv X
ǂv v.
ǂv v. iv, pt. 7
ǂv v. VI
ǂv xii
```

This problem is of course only compounded when the numbering of a series is recorded sometimes in arabic numerals, sometimes in roman.

```
ǂv 2. Reihe, Bd. 6
ǂv I. Reihe, Heft XXII
```

- Numerals[8] are presented in MARC 21 bibliographic records as strings of text characters, not as the values these characters represent. This may lead library systems to arrange series numbers character by character, from left to right, instead of by value.

```
ǂv no. 1
ǂv no. 10
ǂv no. 101
ǂv no. 11
ǂv no. 112
ǂv no. 2
ǂv no. 20
ǂv no. 21
ǂv no. 3
```

Responding to concern over ongoing problems associated with numbered series in automated catalogs, in 1999 the BIBCO Operations Committee, a part of the Program for Cooperative Cataloging (PCC), appointed a Working Group on Series Numbering. The Working Group included these recommendations in its final report, issued the same year:[9]

3. The PCC should contact vendors to work on correcting all numerical sorting of series entries in the OPAC. Support for series sorting (through the entire numbering) is desirable in an integrated library system. Methods already suggested include asking for the input of each vendor's user group as well as having the PCC write a letter to each of the major vendors.

older bibliographic records contain roman numerals in contexts where arabic numerals would be used today.

[8] When used by itself in this report, the word *numeral* refers to the characters 0 (zero) through 9 (ASCII decimal characters 48 through 47; hex characters 30 through 39). The word *numeral* is to be taken to be synonymous with the expressions *digit* and *arabic numeral*. When it is necessary to speak of other types of numbering, the word 'numeral' is always coupled with an adjective: *roman numeral*, *ordinal numeral*.

[9] The text of the report is available at: http://lcweb.loc.gov/catdir/pcc/bibco/seriesnumb.html

4. MARBI and vendors should work towards developing a mechanism which supports disregarding the designation in the series ǂv in its sort of the series.

Portions of the Working Group's report were amplified into Discussion Paper 2001-06, presented to MARBI[10] at its June 2001 meeting in San Francisco.[11] This discussion paper raised the possibility of a coding mechanism for use in subfield ǂv that would allow library systems to arrange series headings in numerical order. Many of those participating in the discussion felt that the problems in the arrangement of series entries stemmed from causes related to library systems, and suggested that solutions should first be sought at the library system level. There was also resistance to the idea of a new coding scheme for series numbering that would need to be applied retrospectively to many millions of bibliographic records. Discussion ended with the suggestion that some division of PCC might design an algorithm to achieve the desired sorting order and share it with library system vendors.

Following the MARBI discussion, PCC's Standing Committee on Automation appointed a task group to develop an algorithm for handling numbered series headings.[12] The task group consisted of Sherman Clarke (New York University), Greta de Groat (Stanford University), Stephen Hearn (University of Minnesota), and Gary L. Strawn (Northwestern University, chair). The task group felt that it was primarily called upon to develop a protocol for normalizing instances of series numbering in bibliographic records. The form of series numbering created by a program following a suggested normalization protocol would be used together with the normalized form of the series heading itself to arrange a set of series entries for display.

The present document, the task group's final report, describes four schemes devised by the task group for the normalization of series numbering. These schemes vary in their sophistication and the correctness of the order of headings they produce. The task group provides algorithms for these four schemes in the hope that they, or other schemes of similar character, will be incorporated into automated library systems and used to display series headings in more useful order.

Although the implementation of a suitable normalization algorithm for series numbering will improve the usability of library catalogs, it will not cure all problems related to series numbering. The letter of transmittal conveying this report contains the task group's observations on areas that appear to call for further examination by other parties. Some of these observations extend to headings, other than series headings, that contain information intended to be arranged in numerical order in library catalogs.

[10] An interdivisional committee of the American Library Association, the Association for Library Collections and Technical Services, the Library and Information Technology Association and the Reference and User Services Association. It is part of the U.S. MARC Advisory Committee.
[11] The text of the Discussion Paper is available at: http://lcweb.loc.gov/marc/marbi/2001/2001-dp06.html
[12] Appendix A contains the charge to the task group.

1.2 Working method

The task group performed all of its work via electronic exchange and the mails. It held no face-to-face meetings.

To obtain a body of series headings against which to test various approaches to the normalization of series numbering, in early March 2002 every numbered series heading in every bibliographic record present in Northwestern University's bibliographic database was extracted and saved to a file. Appendix B describes the characteristics of the series headings in this corpus.

Algorithms for handling series numbering proposed by task group members were translated into program code, and the resulting programs run against the corpus of headings. The results of the tests were inspected and, for the most complicated algorithm, distributed for review and comment. Work proceeded until members of the task group felt that its primary goal had been achieved.

2 Principles underlying the design of the normalization algorithms

2.1 Use in isolation

The normalization techniques described in this report are all designed to be applied to information in one subfield ǂv in a single instance of a series heading in a single bibliographic record when that subfield ǂv is to be considered in the context of the whole heading.[13] The techniques described in this report are to be applied to information in a single bibliographic series heading without consulting any authority record that may support the heading, without considering information in other parts of the same bibliographic record, and without consulting information in other bibliographic records.

The normalization of series numbering involves a set of decisions about the handling of the characters that the numbering contains. The practice generally followed by library systems for normalizing parts of MARC records for use in retrieval and sorting calls for the treatment of each character in isolation; the invariant handling of each character is typically determined by consulting a table. This model cannot always be followed for the normalization of series numbering. In some of the normalization techniques described in this report, each character must be considered as part of a context. A character in an instance of series numbering must at times be regarded in its relation to neighboring characters, and a word must often be regarded in its relation to neighboring words.

2.2 Changes to existing records not called for

In keeping with the general trend of opinion voiced at the MARBI meeting that preceded the formation of the task group, none of the techniques described in this report presumes any change to MARC bibliographic content designation; the normalization techniques

[13] This implies that these techniques are not designed to be applied when generating keyword index terms.

may be applied to information in existing bibliographic records and to records yet uncreated. The techniques reflect an attempt to make the most of the good that resides in bibliographic databases, and to minimize the effect of lower-quality information. Some knotty problems in series numbering can be resolved (at least in the majority of cases) by the application of normalization techniques such as those described in this report; a few can even be resolved in all cases. Some problems (such as those introduced by the presence of roman numerals) remain intractable.

This does not mean that changes to standard practices for recording series numbering would not make the task of sorting series headings simpler and more reliable in the future. The more directly the information in series subfield ╪v can be converted into normalized form, the simpler the task of sorting series headings becomes. The task group identified several important areas in which changes appear to be called for. The task group has included suggestions for changes to practice in the letter of transmittal included with this report.

2.3 The end product of normalization

As a result of the application of the normalization procedures given in this report, the characters that comprise an instance of series numbering are modified in some manner:

- The numeric segments are manipulated so that the series headings to which they are attached may be arranged in numerical order, not in the order dictated by a literal consideration of the characters of which the numbers are composed
- Some alphabetic portions may be removed
- Some alphabetic portions may be converted to a standardized form
- Some alphabetic portions may be converted to normalized form
- Punctuation is replaced by spaces, or removed[14]

The form of the normalized versions of series numbering produced by application of the processes described in this report may give pause to specialists in various fields. It is the arrangement of a set of series headings produced by some normalization algorithm, and not the precise manner in which a series heading is rendered in normalized form by that algorithm, that should be the object of study and criticism. If series headings are arranged in a markedly better manner through the application of one of these algorithms than without it, the appearance of the normalized form should be considered acceptable.

2.4 The best effect with the least damage

A normalization technique should provide for the correct arrangement of as many series headings as possible, while causing as few headings as possible to be arranged in an unsatisfactory manner. The task group recognizes that, absent changes to practice for recording series numbering, no algorithm for achieving the arrangement of series headings can be perfect. The various algorithms given in this report reflect a series of compromises among competing factors: complexity of the algorithm, number of correctly

[14] Before its replacement or removal, punctuation may have a role to play in the normalization of a heading.

sorted series headings, and number of new mislocations. When possible, steps may be taken to avoid the creation of new mislocations.

2.5 Application in various contexts

The techniques described in this report may be applied in various contexts. Among these contexts are the following:

- The techniques may be applied to a set of headings at the time records are stored in a database, and the results stored as index entries for future use
- The techniques may be applied to a set of headings at the time a display is being prepared, and the results used immediately to order the headings in the display

2.6 Normalized forms optimized for sorting

The tasks performed by library systems commonly call for the use of normalized forms of headings. Unfortunately, normalized headings designed for use in searching do not always arrange headings in the desired order; and normalized headings that produce well-ordered displays are often unsuitable for searching.[15] Because search and display operations place competing needs on normalized headings, a single normalized form that works well in both situations probably does not exist. Most library systems nevertheless generate only one normalized form for each access point, and use that normalized form for both indexing and sorting.[16] This single normalized heading is generally optimized for use in retrieval, not for sorting.

The normalized forms of series numbering described in this paper are, by contrast, designed for sorting, not retrieval. The consequences of this design choice are clear. If the normalized series numbering (generated by one of the techniques described in this report) is stored (in a string index) as part of a library system's sole normalized form of a series heading, it will in most cases affect a user's ability to submit a left-anchored search term including a series number and correctly retrieve a single member of a series.[17] If it is felt desirable to provide users with the ability to retrieve individual members of numbered

[15] For example, the NACO normalization standard calls for the retention of the first comma in personal names that consist of a surname and forename. This comma serves to different between names that contain the same alphabetic characters but yet do not represent the same person ('Xiao Lan' vs. 'Xiao, Lan'). However, the use in searching of the normalized form with the comma would cause a problem—catalog users would have to remember to include the comma in search terms.

[16] This normalized form is typically generated at the time a new or modified record is stored. Some library systems generate more than one normalized index entry from some access points. For example, some generate both a name/title and a title access point from a name/title heading. But the important points remain the same: the normalized index entries are generated ahead of time, and are used for both retrieval and sorting.

[17] One library system examined during the preparation of this report allows the user to input a series heading plus the numbering to retrieve a single member of a series; the series heading plus numbering supplied by the user must of course exactly match the series information in the bibliographic record. The use of one of the normalization techniques described in this document by systems that use a single index entry for both retrieval and sorting could therefore disable an existing system feature in some cases, although one perhaps not used very often.

series, the capabilities of local systems that employ a single normalized form for both retrieval and sorting will need to be enhanced.[18]

2.7 Areas not covered

Even the most complicated of the techniques described in this report is not able to resolve all problems created by changes to numbering patterns introduced by publishers, and all inconsistencies in numbering created by catalogers. The algorithms are able deterministically to produce the normalized form for each subfield ǂv, but the order of series headings produced by the action of an algorithm may always not be the 'right' order (and, indeed, in some cases there may be no discernable right order).

Here are some examples of the many variations and other problems for which none of the algorithms proposed in this report attempts to provide a universal solution:[19]

- Change to or variation in the pattern of series numbering

  ```
  ǂv v. 1
  ǂv v. 2
  ǂv v. 3
  ǂv v. 4
  ǂv v. 5
  ǂv 1997/1998, v. 6
  ǂv 1999
  ```

- Infrequently-occurring typographical errors

  `Abhandung` *instead of* `Abhandlung`
  `Vd.` *instead of* `Bd.`

- Designation for a subseries sometimes given as part of the series heading, sometimes as part of the series numbering.

  ```
  Heading ; ǂv 3. Reihe, Bd. 14
  Heading. ǂn 3. Reihe ; ǂv Bd. 15
  ```

- Extraneous text (such as the name of the series editor) included with the series numbering.

  ```
  ǂv 9. (supplement-) th. Elberfeld, 1861
  ```

[18] The reference here is to the retrieval of the members of a series via a search of a left-anchored index. If the library system includes the contents of subfield ǂv in its keyword index, it should already be possible for a clever searcher to retrieve an individual member of series via a keyword search.

[19] The corpus of series headings extracted from the Northwestern University bibliographic database contains these and many other irregularities. Any database of any comparable size, especially one that has seen substantial additions via bulk loads, will contain similar surprises.

- The lowercase letter ‘l’ (‘el’) or the uppercase letter ‘I’ (‘eye’) used instead of the digit ‘1’, or vice-versa.
- The letter ‘O’ (‘oh’) used instead of the numeral zero, or vice-versa.
- Elements within the series numbering given in varying order

```
ǂv 1979, no. 1
ǂv no. 1, 1980

ǂv no. 40, 98th Congress
ǂv 98th Congress, no. 31

ǂv no. 53, 3d series
ǂv 2nd ser., no. 31
```

- Year recorded with only two digits.[20]
- Numbers (either cardinal or ordinal) presented as words.
- Decimal numbers[21]

[20] In an attempt to find a solution to the problem of years represented by two digits, especially in an effort to cause entries for items published in 2000 and later years to file after entries for 1999 and earlier years, the task group performed an experiment in which a program expanded a sequence of two consecutive digits to four when the two digits in question matched the last two digits of the date of publication (008 field, bytes 9-10). Even though the algorithm considered the date of publication, expansion of two digits to four in this manner was found to be unreliable, and caused more dislocations than existed when the putative two-digit years were left untouched.

Example: *Aircraft accident incident summary report.* (Except in the matter of the two-digit year, the series numbering is normalized here in the form described in section 3.4 or 3.5.)

Numbering	Year of publication	Normalized form
NTSB/AAR-87/04/SUM	1988	NTSB AAR 00000087 00000004 SUM
NTSB/AAR-87/03/SUM	1987	NTSB AAR 00001987 00000003 SUM

Because NTSB/AAR-87/03/SUM was published in 1987, the ‘87’ in its series numbering can (correctly) be converted to ‘1987’. But because NTSB/AAR-87/04/SUM was published in 1988, its ‘87’ cannot reliably be converted to ‘1987.’ If putative two-digit dates were expanded to four digits by comparing them to the date of publication, 87/03 (i.e., 1987/03) would incorrectly come *after* 87/04 (as shown above)—possibly a good deal after; but if the suspected dates were left alone, the headings would be in the correct order.

Many two-digit numbers that accidentally match the year of publication would also mistakenly be converted by an attempt to fix two-digit years. For example, the numbering ‘91-91B’ in a bibliographic record for an item that happened to be published in 1991 might improperly be normalized as ‘00001991 00000091 B’.

[21] The corpus contains 2990 headings whose series numbering includes a full stop between two digits—one signature of a decimal number. In its review of these headings, the task group determined that while some of these were decimals numbers, most represent ‘segmented’ whole numbers. (For example, the members of one series bore these numbers: 80.3, 80.6, 80.10 and 80.11, and the members of another series bore these numbers: 4000.8, 4000.9, 4000.10, 4000.11 and 4000.12; these were deemed to be segmented numbers, not decimals. One series that clearly does use decimal numbering is *Blue suede shoes*, which instituted a

- Numbers presented as roman numerals
- Months and seasons represented by words instead of numbers; months and seasons appearing before of the year

All this notwithstanding, some of the provisions of some of the techniques described in this report are intended to overcome variations in series numbering—variations caused by changes in practice over approximately one hundred years, inconsistency on the part of publishers, and human error.[22]

3 Normalization algorithms

3.1 Introduction

The task group identified several approaches to the normalization of series numbering. The following divisions of section 3 of this report describe each technique in detail sufficient to allow a library system vendor to implement the technique successfully. In most cases, the description of an algorithm takes only a few lines. Because of its complexity, substantial detail must be included in the description of level 4 normalization.

These instructions use the verb *remove* to indicate that one or more characters should disappear and the adjacent characters should be drawn together; this verb does *not* mean that the character or characters are replaced by a space.

Except when explicitly stated otherwise, the noun *word* refers to a contiguous series of characters bounded by blank spaces.

It should be assumed throughout that leading and trailing spaces are always removed, and that multiple occurrences of a space are always reduced to a single space.

The programs used for testing these algorithms employed ‘NACO’ normalization whenever system normalization was called for. NACO normalization is described at: http://www.loc.gov/catdir/pcc/naco/normrule.html.

The lists of words mentioned at several points in the following sections are contained in Appendix D.

system of random decimal numbers after the first 15 volumes.) There is no way for a program to determine whether a number such as ‘1.2’ represents a decimal fraction or two whole numbers—especially without inspecting the numbering of other members of the series. The task group felt that algorithms should be designed correctly to handle the majority of cases, and should therefore treat the full stop between two digits as marking the segments of a number.

[22] A few cases in which allowance can safely be made for operator variation are indicated in the report. There is no reason that developers of library systems could not extend the techniques described in this report to account for other similar conditions that may be found.

3.2 Level 1: Justify numerals to standard length

Technique:

Manipulate each contiguous sequence of digits (delimited by any non-digit) as described in Appendix C.

Perform standard normalization on the remainder of the series numbering subfield.

The accompanying table shows selected series numbering from a few different headings as normalized by this technique.

Series numbering in subfield ǂv	**Becomes**
v. 5	V 00000005
1997/1998	00001997 00001998
no. 53, 3d series	NO 00000053 00000003 D SERIES
no. (PHS) 79-3216	NO PHS 00000079 00003216

This simple technique produces a marked improvement in the arrangement of series headings, and introduces no new mislocations. Indeed, because this technique is so simple and entails no mislocation of headings that are not already disordered, it is difficulty to understand why at least this level of normalization has not been widely adopted by library system vendors.

The series numberings in each of the following tables are extracted from the members of a single series, which are correctly arranged by this algorithm in ascending order.

Series numbering in subfield ǂv	**Becomes**
1	00000001
2	00000002
10	00000010
23	00000023
129	00000129

Series numbering in subfield ǂv	**Becomes**
Bd. 2	BD 00000002
Bd. 9	BD 00000009
Bd. 10	BD 00000010

Series numbering in subfield ǂv	**Becomes**
1950 no. 1	00001950 NO 00000001
1950, no.2	00001950 NO 00000002
1950 no. 10	00001950 NO 00000010

This technique fails to address the problem of variant captions and variant subseries designations. The series numberings in each of the following tables are extracted from the members of a single series. These headings are not arranged correctly by this algorithm.

Series numbering in subfield ǂv	Becomes
13	00000013
v. 9	V 00000009
vol. 3	VOL 00000003

Series numbering in subfield ǂv	Becomes
Bd. 54	BD 00000054
Heft 37	HEFT 00000037

Series numbering in subfield ǂv	Becomes
2a s´erie, 3	00000002 A SERIE 00000003
2. s´er., 1	00000002 SER 00000001
8	00000008
s´er. 2, 2	SER 00000002 00000002

3.3 Level 2: Justify numerals; remove prefix

Technique:

```
If the numbering contains any digits
    Discard all characters that precede the first digit
    Manipulate each contiguous sequence of digits as described in Appendix C
Perform standard normalization on the remainder of the series numbering subfield
```

The accompanying table shows selected series numbering from a few different headings as normalized by this technique.

Series numbering in subfield ǂv	Becomes
v. 5	00000005
1997/1998	00001997 00001998
No. 53, 3d series	00000053 00000003 D SERIES
No. (PHS) 79-3216	00000079 00003216

This brute-force approach to text that may represent a caption produces surprisingly good results in many cases. The series numberings in the following tables are each extracted from the members of a single series, which are correctly arranged by this algorithm in ascending order.

Series numbering in subfield ǂv	Becomes
57	00000057
66	00000066
no. 74	00000074
75	00000075
no. 81	00000081
v. 86	00000086
no. 92	00000092

Series numbering in subfield ǂv	Becomes
1. Bd.	00000001 BD
Bd.5	00000005

Unfortunately, the technique creates mislocations when at least a portion of the alphabetic information that precedes the first digit is significant; and it fails to solve problems caused by internal alphabetic characters that vary from heading to heading. The mislocations produced by this method can be especially troubling when a series is divided into two or more subseries: the entries for the various subseries are often intermingled. The following tables show series numberings not correctly handled by this technique. Each table contains numbering from several instances of a single series heading.

Series numbering in subfield ǂv	Becomes
nova ser., 1	00000001
2. ser., 5	00000002 SER 00000005
3	00000003
n.s., 4	00000004
5	00000005

Series numbering in subfield ǂv	Becomes
2	00000002
2. ser., 25	00000002 SER 00000025
3	00000003
3. ser., 3	00000003 SER 00000003
3. ser., 25	00000003 SER 00000025
6	00000006
6. ser., 16	00000006 SER 00000016
7a ser., 25	00000007 A SER 00000025
7. ser., 2	00000007 SER 00000002
8. ser., 4-	00000008 SER 00000004
13	00000013
20	00000020

Series numbering in subfield ǂv	Becomes
CH-18	00000018
NS-33	00000033
NP-75	00000075

Series numbering in subfield ǂv	Becomes
92d Congress, no. 14	00000092 D CONGRESS NO 00000014
92nd Congress, no. 3	00000092 ND CONGRESS NO 00000003

Series numbering in subfield ǂv	Becomes
vol. IV, no. 2	00000002
vol. 3, no. 6	00000003 NO 00000006

Series numbering in subfield ǂv	Becomes
level 1	00000001
prebasic materials, level 2	00000002
basic materials, level 2	00000002
prebasic materials, level 2[23]	00000002
basic materials, level 2	00000002
basic materials, level 3	00000003
prebasic materials, level 4	00000004
basic materials, level 4	00000004
prebasic materials, level 5	00000005

3.4 Level 3: Justify numerals; remove prefixes that appear to be captions

Technique:

```
Manipulate each contiguous sequence of digits as described in Appendix C
Perform standard normalization on the remainder of the series numbering subfield[24]
For each word in the normalized numbering, from left to right
    If the word is present in a list of caption words (Appendix D, both single-character and multi-
            character captions, and also including uppercase letter 'V' and the 'number sign'[25])
        Remove the word and continue with the next word
    Else (the word is not present in a list of caption words—including words composed of digits)
        Accept the remainder of the series numbering as normalized
```

The accompanying table shows selected series numbering from a few different headings as normalized by this technique.

Series numbering in subfield ǂv	Becomes
v. 5	00000005
1997/1998	00001997 00001998
No. 53, 3d series	00000053 00000003 D SERIES
No. (PHS) 79-3216	PHS 00000079 00003216

[23] It might be argued that the arrangement by 'level' achieved by the level 2 algorithm is preferable to a sort that keeps together all similar items (all 'prebasic materials', for example). But, because the caption is disregarded, the technique is not able to collocate multiple instances of the same series number (for example, 'basic materials, level 2' appears at two different points); this must be regarded as an error.

[24] Up to this point, this technique is identical with the level 1 technique.

[25] MARC character 35, hex 23. (All of the values given here for MARC characters refer to 8-bit encoding.)

Like the technique described in section 3.3, this technique produces correct sorting for many series, but like that technique it fails to address some problems, and creates mislocations into the bargain. Among the problems not addressed are alphabetic prefixes that correspond to the title of the series, ordinal numbers, designations for subseries, and internal captions. Among the causes for mislocations created by this routine are the occasional removal of significant text (for example, the roman numeral 'V') that happens to match caption text. The following tables show series numberings not correctly handled by this technique. Each table contains numbering from several instances of a single series heading.

Series numbering in subfield ǂv **(series heading: 'Advisory circular')**	**Becomes**
70/7460-2J	00000070 00007460 00000002 J
90-45A	00000090 00000045 A
AC 70/7460-2F	AC 00000070 00007460 00000002 F
AC 90-428	AC 00000090 00000428

Series numbering in subfield ǂv	**Becomes**
ser. 1a. no. 16	00000001 A NO 00000016
ser. 1, no. 9	00000001 NO 00000009

Series numbering in subfield ǂv	**Becomes**
ser. F, v. 5	00000005
ser. D, v. 7	00000007

Series numbering in subfield ǂv	**Becomes**
T 141	00000141
C 151	C 00000151

Series numbering in subfield ǂv	**Becomes**
Band 2, Heft 5	00000002 HEFT 00000005
Bd. 2, hft. 2	00000002 HFT 00000002
2. Reihe, Bd. 2	00000002 REIHE BD 00000002
Bd. V, Heft 3	00000003
Bd. 3, hft. 1	00000003 HFT 00000001

Series numbering in subfield ǂv	**Becomes**
salle n. 9	SALLE N 00000009
salle no 6	SALLE NO 00000006

Series numbering in subfield ǂv	**Becomes**
6, Bd. 6	00000006 BD 00000006
6, pt. 2	00000006 PT 00000002

Series numbering in subfield ╪v	**Becomes**
no. (SSA) 85-001	SSA 00000085 00000001
no. (SSA) no. 13-11746	SSA NO 00000013 00011746

Series numbering in subfield ╪v	**Becomes**
new ser. 2	NEW SER 00000002
new ser., 39	NEW SER 00000039
new series, 3	NEW SERIES 00000003

Series numbering in subfield ╪v	**Becomes**
n.s., 2	00000002
3	00000003
n.s. 10	00000010
n.s. 24	00000024
nova ser. 27	NOVA SER 00000027
Nova series 18	NOVA SERIES 00000018

3.5 Level 4: Elaborate normalization

3.5.1 Introduction

This algorithm, by far the most elaborate of the four proposed by the task group, attempts to eliminate the drawbacks present in the other algorithms. This algorithm regularizes numbering for subseries (both numbered and not numbered), removes internal captions, converts ordinal numbers into digits, and solves other problems not addressed by the other three techniques. This algorithm attempts to avoid the mislocations created by some of the other techniques.

To achieve these objectives, it was found necessary to design one method of handling for series numbering that contains at least one digit, and a separate method for series numbering that contains no digits. The reason for this difference in treatment is that series numbering without digits presents an especially large number of difficult problems, starting with but not limited to problems in the identification of captions; while numbering with digits contains landmarks that can guide a normalization routine to a correct result in most cases. (For similar reasons, the algorithm treats alphabetic characters that follow the last digit in a manner different from that employed for alphabetic characters that precede the final digit.) The scheme proposed for series numbering that contains digits is fairly aggressive in its handling of text; but the scheme for series numbering without digits is more conservative.

Some parts of this algorithm reduced quite nicely into a few statements; other parts could not be reduced so handily, and will read very much like program code. In fact, some parts of the algorithm are nearly literal translations of the code used to perform these actions. In these cases, it seemed impossible to describe the intricate work being performed in any other manner.

Apply the instructions in section 3.5.2 to all series numbering. Then apply the instructions in either section 3.5.3 (series numbering that does not contain digits) or section 3.5.4 (series numbering that contains digits).

The following table shows the application of this algorithm to selected instances of series numbering. For additional examples, see sections 3.5.3.4 and 3.5.4.6.

Series numbering in subfield ǂv	**Becomes**
v. 5	00000005
1997/1998	00001997 00001998
no. 53, 3d series	00000053 SER 00000003
no. (PHS) 79-3216	PHS 00000079 00003216
Nouv. sér., B	NEW SER B
Book A	BK A
OEA/SER.H/XII (English)	OEA SER H XII ENGLISH
levels C-F	LEVEL C F

The following tables show series numberings correctly handled by this technique. Each table contains numbering from several instances of a single series heading.

Series numbering in subfield ǂv	**Becomes**
3	00000003
5	00000005
nova ser., 1	NEW SER 00000001
n.s., 4	NEW SER 00000004
2. ser, 5	SER 00000002 00000005

Series numbering in subfield ǂv	**Becomes**
no. 109	00000109
no. 120	00000120
publication no. 158	00000158
no. 161	00000161

Series numbering in subfield ǂv	**Becomes**
n. s., no 1, etc.	NEW SER 00000001 ETC
nouv. sér., no 4	NEW SER 00000004
n. s., no 6.	NEW SER 00000006
nouv. sér., no 15	NEW SER 00000015
n. s., no 29	NEW SER 00000029

Series numbering in subfield ǂv	**Becomes**
no. 7-1-2	00000007 00000001 00000002
no. 7-3AB	00000007 00000003 AB
no. 7-9A	00000007 00000009 A
no. 7-10A	00000007 00000010 A

no. 7-11	00000007 00000011

Series numbering in subfield ǂv	Becomes
Bd. 1, 1. T.	00000001 00000001
Bd. 1, Teil 2	00000001 00000002
Bd.3	00000003

Series numbering in subfield ǂv[26]	Becomes
no. 25	00000025
JP 26.	00000026
no. 27	00000027

Series numbering in subfield ǂv	Becomes
v. 3, supplement c	00000003 SUP C
v. 13, suppl. A	00000013 SUP A
v. 13, suppl. B	00000013 SUP B

Series numbering in subfield ǂv	Becomes
ser. 2a, 17	SER 00000002 00000017
ser. 2, 20	SER 00000002 00000020
ser. 2a, 24	SER 00000002 00000024

Series numbering in subfield ǂv	Becomes
92nd Congress, no. 3	00000092 CONG 00000003
92d Congress, no. 14	00000092 CONG 00000014

At several points during this work, the instruction is given to replace one or more characters with a *caption substitute*. The caption substitute may be any sequence of characters not likely otherwise to appear within a subfield. The caption substitute is inserted into series numbering when a character or characters are recognized as representing a caption. It is simply a placeholder, and serves at a later stage to identify the place from which a caption was removed. (The caption substitute eventually disappears from the normalized numbering.) The program used for testing the level 4 algorithm used a word consisting of a space, the MARC delimiter and a second space as the caption substitute.[27] (Since the end of the series numbering is identified by the start of the next subfield—if any—the series numbering cannot by definition contain a MARC delimiter.) The illustrations in this report show a vertical bar surrounded by spaces as the caption substitute. The caption substitute is only relevant in series numbering that contains digits, but it may also be used in other numbering.

[26] The title of the series, *Joint publication (Great Britain. Historical Manuscripts Commission)* causes the algorithm to treat 'JP' as caption information.

[27] The spaces around the caption substitute guarantee that it will be treated at a later point as a separate word.

3.5.2 Initial steps

3.5.2.1 Introduction

Except as instructed, apply the instructions in the various subdivisions of section 3.5.2 to all instances of series numbering, without regard to the presence or absence of digits.

The following examples show series numbering modified by the instructions contained in section 3.5.2. This intermediate form of the series numbering will be further modified by additional operations, described in sections 3.5.3 and 3.5.4.

Series numbering in subfield ǂv	Becomes
4	4
Bd. 1.	BD 1
No. 24744-24747	NO 24744 24747
fiche 4,293-4,296[28]	FICHE 4,293 4,296
91-14S	91 14S
DS5-S1	DS5S1
30, etc.	30, ETC
1981 no. 10	1981 NO 10
DP 97-002	DP 97 002
year one, vol. D.	YEAR ONE, VOL D
no. (ADM)	NO ; ADM
OEA/Ser.H/XII (English)	OEA/SER H/XII ; ENGLISH
v. III	V III
lre ptie.	LRE PTIE
ch. R, S	R, S

3.5.2.2 Remove certain 'chapter' abbreviations

These instructions only apply to series numbering that does not contain digits.[29]

```
Remove all punctuation that occurs before the first non-punctuation character[30]
If the initial characters in the series numbering match characters that are abbreviations for
            'chapter'
    Remove those characters[31]
Else if the initial characters in the series numbering (in any mixture of uppercase and lowercase
            letters[32]) match any words that mean 'chapter'
    Remove those characters
```

[28] Note in this example that it is not possible easily to determine whether the series consists of fiche 4, fiche 293 through 294 and fiche 296; or of fiche 4293 through 4296.

[29] Similar operations are performed as part of the routine work for numbering that contains digits.

[30] The principal punctuation character of interest at this point is the opening square bracket; but any leading punctuation should also be removed.

[31] There is no need to use the caption substitute, because the numbering contains no digits. There would be no harm if the caption substitute were inserted here; it will simply be removed later.

[32] At this point, lowercase characters *must not* yet have been converted to uppercase.

3.5.2.3 Replace unusual abbreviations for ‘number’

These instructions apply to all instances of series numbering.

This section describes an attempt to adjust the series numbering for several occasionally-encountered abbreviations for ‘number’, such as:

> n′ (*where the ' is the soft sign*)[33]
> n^{o} (*where the* ^{o} *is the degree sign,*[34] *superscript zero,*[35] *or angstrom*[36])[37]

```
Scan through the series numbering from right to left.[38]
If the current character is a superscript zero, a soft sign, an angstrom, or a degree sign:
    If the current character is the first character in the series numbering
        Replace the current character with its standard normalized form[39]
    Else (the current character is not the first character in the numbering)
        If the character preceding the current character is ‘N’ or ‘n’
            Replace the current character and the preceding character with the caption substitute
        Else if the preceding character is a full stop or a colon
            If the current character is the second character in the numbering
                Replace the current character with its standard normalized form[40]
            Else (the current character is the third or later character in the numbering)
                If the character preceding the full stop or colon is ‘N’ or ‘n’
                    Replace that character, the full stop or colon and the current character with
                            the caption substitute[41]
                Else (the character preceding the full stop or colon is not ‘N’ or ‘n’)
                    Replace the current character with its standard normalized form[42]
        Else if the character preceding the current character is the numeral ‘8’
            If the current character is a superscript zero
                Remove the current character[43]
```

[33] MARC character 167, hex A7. The corpus of series headings contains 17 uses of the soft sign in this manner.

[34] MARC charcter 192, hex C0.

[35] The representation of the superscript zero in the MARC 21 communications format is somewhat difficult to explain in a few words. Consult *MARC 21 specifications for record structure, character sets, and exchange media* (2000), p. 24-26. Note in any case that the MARC specifications describe records exchanged between systems; the superscript characters, and indeed all of the characters, may be represented within library systems in some other manner.

[36] MARC character 234, hex EA.

[37] The corpus of series headings contains 6 uses of the superscript zero and 8 uses of the angstrom in this manner. While the corpus contains no uses of the degree sign in this manner, it was felt best to include that character here because of its similarity in appearance to the superscript zero, and the likelihood that it might be misused in a similar manner.

[38] Here and elsewhere, scanning a string from right to left is described, because the work under consideration entails a change to the length of a string; scanning from right to left allows a routine to determine the length of the string only once. There appears to be no particular reason the same work could not be performed in a left-to-right fashion, if such an approach were able efficiently to adapt to the length of the string as it changes.

[39] Example: ‘°arrskrift 62’ becomes ‘arrskrift 62’.

[40] The corpus of series headings contains no example of this condition.

[41] Example: ‘n.° 106, anexo 5’ becomes ‘ | 106, anexo 5’.

[42] The corpus of series headings contains no example of this condition.

```
Else (preceding character is not ‘N’, ‘n’, full stop, colon or ‘8’)
    If the current character is not a superscript zero
        Replace the current character with its standard normalized form[44]
    Else (the current character is a superscript zero)
        Do nothing at this time. Superscript characters will be dealt with later as a class,
                and it is critical that this character not be translated until its context is
                examined in more detail
```

Series numbering	**Becomes**[45]
with superscript zero: n^{0} 55	| 55
with superscript zero: Collection in -8^{0}, t. 21, fasc. 3	Collection in -8, t. 21, fasc. 3
with angstrom: VIII. bd., n° III.	VIII. bd., | III.
with angstrom: 29. °arg, nr. 3	29. arg, nr. 3
with angstrom: n.° 3	| 3
with angstrom: [1. s´er. (in 8°)] fasc. 159	[1. s´er. (in 8)] fasc. 159
with soft sign: n.´ 30	| 30
with soft sign: Volume n´ 70	Volume | 70
with soft sign: v´ip. 27	vip. 27

3.5.2.4 Replace abbreviations for ‘number’ that contain internal punctuation

These instructions apply to all instances of series numbering.

Replace each occurrence of the abbreviations ‘N.O’, ‘N:O’, ‘N;O’, ‘N:0’[46] and ‘N:R’[47] (in any mixture of uppercase and lowercase letters) with the caption substitue.[48]

[43] The abbreviation 8^{0} for *octavo* occurs 2 times within the series numbering in the corpus of series headings. (It occurs with far greater frequency in the series headings themselves.) Other corpora may contain similar abbreviations that include superscript zero for *quarto* and other publication formats; if so, the technique described here should carefully be expanded to include them.

[44] Examples: ‘pis´mo 16’ becomes ‘pismo 16’; ‘29. °arg., nr. 3’ becomes ’29. arg., nr. 3’.

[45] These examples show the numbering as normalized only up to this point.

[46] The oval character in this abbreviation is a zero; the similar character in the other abbreviations is the uppercase letter ‘O’ (‘oh’).

[47] AACR2 Appendix B.9 specifies the abbreviation ‘n:o’ for the word ‘numero’ in Finnish, and does not authorize the other abbreviations. The corpus of series headings contains 13 instances of ‘n:r’, one instance of ‘n;r’, 20 instances of ‘n.o’, one instance of ‘n:0’ (with a zero) and 142 instances of ‘n:o’.

[48] The original implementation of the routine that performed this work included an elaborate series of tests designed to guarantee that the putative abbreviation with internal punctuation was in fact a separate word. On further inspection, it was discovered that the corpus of series headings contained no instance of any of these sequences of characters that did *not* pass all these complicated tests—i.e., it contained no instance of one of these sequences that didn’t end up getting replaced with the standard abbreviation. Therefore, the much simpler algorithm described here—find the characters and replace them—was substituted. Should an instance of series numbering containing one of these abbreviations that should *not* be replaced ever be encountered, this algorithm will have to be made more elaborate.

Series numbering	Becomes[49]
N:o 96.	| 96.
6. följden, ser. B, bd. 6, n:o 3.	6. följden, ser. B, bd. 6, | 3.
N:r 371	| 371
n.o 9]	| 9]
n:o 496. Arsbok 42 (1948) n:o 5	| 496. Arsbok 42 (1948) | 5

3.5.2.5 Replace single-character abbreviations

These instructions only apply to series numbering that contains digits. It is only necessary to perform this work if the series numbering contains 2 or more characters.

Caption information is often abbreviated to a single letter. Unfortunately, the same single letters may also represent non-caption information. A careful attempt must be made to distinguish between the two cases, so that single-letter caption information can safely be deleted, and single-letter non-caption information retained. Perhaps only because the letter ‘v’ appears alone more often than the other letters, in a larger number of guises, and is also commonly used as a roman numeral, it has been found expedient to handle the letter in one manner, and to handle all other potential single-letter abbreviations in another manner.[50]

Letters ‘d’, ‘f’, ‘h’, ‘j’, ‘n’, ‘p’, ‘r’, ‘s’, ‘t’ and their uppercase equivalents

Perform the work described here for the letter in its uppercase form, and separately for the letter in its lowercase form. Unless explicitly instructed otherwise, retain each character as found.

```
Scan the series numbering from right to left
For each occurrence of the character of interest that occurs to the left of the rightmost numeral[51]
    If the current character is ‘f’, ‘F’, ‘p’, ‘P’, ‘s’ or ‘S’ (i.e., the character is a potential abbreviation
            for ‘series’, and so may be part of an abbreviation for ‘new series’)
```

[49] These examples show the numbering only as normalized up to this point.

[50] In an experiment, the letter ‘v’ was processed by the algorithm used for other single-letter captions. The experiment revealed 86 cases in which the handling differed from that provided by the separate algorithm. Sixteen of these differences were found to be corrections to errors made by the special algorithm, the remainder were new errors made by the common algorithm. (For example, the ‘v’ was not recognized by the common algorithm as a caption in the following instances of series numbering: ‘v219’, ‘6. följden, ser.B v.1, no. 8’, and the second ‘v’ in ‘v. 31, no. 2-v. 32’.) While it would be possible to add special coding for these exceptional occurrences of ‘v’ to the common algorithm, to do so would be in effect to recreate the separate algorithm within the common algorithm, to no obvious advantage. Instead, an attempt was made to change the separate algorithm for ‘v’ to handle additional cases correctly. It is interesting to note also that the use of the common algorithm instead of the special algorithm for ‘v’ reduced the time required to process the corpus of headings by about 10%. This indicates thast optimization of the special routine for ‘v’ would be well repaid.

[51] Retain at this point all characters that occur to the right of the rightmost numeral.

```
    If the following word[52] is in the list of ‘new’ words or if the preceding word[53] is in the list of ‘new’ words and the word before that word (if any) is not in the list of  ‘series’ words
        Retain the current character for now[54]
    Else if the current character is ‘s’ or ‘S’
        If the following word is in the list of ‘hors’ words or if the preceding word is in the list of ‘hors’ words and the word before that word (if any) is not in the list of ‘series’ words
            Retain the current character for now
    Proceed as directed below
Else if the character is ‘h’ or ‘H’ (i.e., the character is a potential abbreviation for ‘hors’, and so may be part of an abbreviation for ‘hors series’)
    If the following word is in the list of ‘series’ words or if the preceding word is in the list of ‘series’ words and the word before that word (if any) is not in the list of  ‘hors’ words
        Retain the current character for now
    Proceed as directed below
Else if the character is ‘n’ or ‘N’ (i.e., the character is a potential abbreviation for ‘new’, and so may be part of an abbreviation for ‘new series’)
    If the following word is in the list of ‘series’ words or if the preceding word is in the list of ‘series’ words and the word before that word (if any) is not in the list of  ‘new’ words
        Retain the current character for now[55]
    Proceed as directed below
Else (current character is none of those mentioned above)
    Proceed as directed below
```

If the preceding instructions say ‘proceed as directed below’

```
If the current character is the first character in the series numbering
    Apply the Standard handling described below
Else[56]
```

[52] Here and elsewhere *in this section*, isolate the ‘following word’ in this manner. Temporarily apply standard normalization to the portion of the series numbering that begins at the *second* character following the character of interest, skipping over one intervening character, whatever it might be. Isolate from this temporarily normalized heading the first word. Isolating the word in this manner and then comparing it to a list of words effectively determines (in a roundabout way) that the character of interest and the following word are actually separate words.

[53] Here and elsewhere *in this section*, isolate the ‘preceding word’ in this manner. If the character preceding the character of interest (disregarding any intervening spaces) is a comma, the character of interest is part of a sequence or is the beginning of a new level of hierarchy in the numbering; the system should behave as if the preceding word was not found in any of the lists mentioned here. Otherwise, temporarily normalize the series numbering from its beginning up to but not including the character of interest, and isolate from this the final word.

[54] The instruction ‘Retain the current character’ means: leave the current character as found, and continue with the next character in the series numbering. Examples: In the numbering ‘new series, v. 21.’, the first ‘s’ meets this test; it is (for the purposes of the work described) here retained in the numbering because it is preceded by a word that means ‘new’. The other ‘s’ in this numbering is made subject to other tests described here. (The second ‘s’ will also be retained, but for other reasons.) In the numbering ‘n.F., Heft 38’, the ‘F’ passes this test, and is retained at this point; but the ‘t’ is made subject to other tests described here. This test has occasional surprising side effects. For example, this test allows the ‘H’ in the series numbering ‘no. (PHS) 91-1768’ to be retained (because it is preceded by ‘P’, which is a ‘series’ word). (Obviously, the ‘H’ should be retained in any case.)

[55] Example: in the numbering ‘n.F., Heft 38’, the ‘n’ passes this test, and is retained at this point.

```
If the character preceding the current character is an opening bracket, opening
        parenthesis, opening brace or greater than sign
    If the character following the current character is a colon, comma, full stop, closing
            bracket, closing parenthesis, closing brace or less than sign
        Apply *Standard handling*[57]
    Else
        Retain the current character[58]
Else
    Apply *Standard handling*[59]
```

Standard handling:

```
If the character following the current character is a space
    If the current character is the uppercase form of a letter
        Retain the current character[60]
    Else (current character is the lowercase form)
        Replace the current character with the caption substitute[61]
Else if the character following the current character is a letter of the alphabet
    Retain the current character[62]
Else if the character following the current character is a colon or comma
    If the current character is the uppercase form
        Retain the current character[63]
    Else (current character is the lowercase form)
        If the next significant character following the colon or comma (disregarding any
                intervening spaces or punctuation) is a digit
            Replace the character with the caption substitute[64]
        Else (next significant character is not a digit)
            Temporarily apply standard normalization to the portion of the series
                    numbering following the colon or comma, and isolate from this the
                    first word
```

[56] The character of interest cannot be *at the end* of the series numbering, as this whole body of instructions is applied only to characters that appear to the left of the last numeral in the series numbering.

[57] Example: apply standard handling to the 'T' in 'Bd. 4, Lft. N (T.1)'; apply standard handling to the 's' in '21-22, 1984-85 (s.108-215)'; apply standard handling to the 't' in '[t. 14]'.

[58] Example: retain the 'n' in the numbering 'no.7'.

[59] Examples: retain the 'H' in 'n.F., Heft 33'; retain the 's' and both appearances of 'n' in '73, section C, no. 7'.

[60] Example: in the numbering 'Bd. 4, Lfg. N (T.1), the 'N' is retained. (The 'T' will be replaced by the caption substitute.)

[61] Example: the numbering 't 3,7' becomes '| 3,7'; the numbering 'n 94-95' becomes '| 94-95'. The corpus of series headings contains no examples of one of these lowercase single-character abbreviations followed by a space that is *not* followed by a numeral; in other words, there is no need for special testing—all may simply be removed. Should the need for such a test be discovered later, the following (from an earlier version of this algorithm) is suggested:

```
If the character following the space (disregarding punctuation and additional spaces) is a numeral
    Replace the current character with the caption substitute
If the character following the space in the non-normalized numbering is a full stop
    Replace the current character with the caption substitute
Else
    If the first word in the normalized portion of the numbering following the space is *not* in the list of
            caption words
        Replace the current character with the caption substitute
```

[62] Examples: retain the 'f' in 'fasc. 58'; retain the 'n' in 'no. 71'.

[63] Example: retain the 'D' in 'D:8'.

[64] Example: the numbering 'n:3' becomes '| :3'; the numbering 't, 56, etc.' becomes '| , 56, etc.'

```
            If this isolated word is present either in the list of multi-character caption
                    words or in the list of single-letter caption words
                Replace the current character with the caption substitute
            Else (isolated word does not appear to be a caption word)
                Retain the current character
Else if the character following the current character is a full stop
    If the next significant character following the full stop (disregarding any intervening
            spaces or punctuation) is a digit
        If the current character is the third or later character in the numbering
            If the character preceding the current character (disregarding any intervening
                    space) is a full stop and if the character immediately preceding the
                    full stop is an uppercase character
                Retain the current character[65]
        Replace the current character with the caption substitute
    If the current character is the uppercase form (i.e., the character may be part of an
            initialism)
        If the character following the full stop (disregarding any intervening spaces) is an
                uppercase letter
            If this following uppercase character is itself followed by a full stop
                Retain the current character[66]
            Else
                Replace the current character with the caption substitute[67]
        Else if the character following the full stop is a numeral that is itself followed by a
                full stop and if the current character is ‘F’, ‘P’ or ‘S’
            Replace the current character with the series substitute[68]
        Else if the character preceding the character of interest (disregarding any
                intervening spaces) is a full stop
            If the character preceding this full stop is an uppercase letter
                Retain the current character[69]
            Else if the character preceding this full stop is a digit and the current
                    character is ‘F’, ‘P’ or ‘S’
                Replace the current character with the series substitute[70]
            Temporarily normalize the portion of the series numbering beginning with the
                    second character after the current character
            If the first word in the normalized remainder of the series numbering is not
                    present in either the list of single-character caption words or the list
                    of multiple-character caption words
                If the current character is not the first character in the series numbering
                    Temporarily normalize the portion of the series numbering that
                            precedes the current character
                    If the last word in this normalized portion of the series numbering is
                            not present in the list of multiple-character caption words
                        Replace the current character with the caption substitute
            Else (the next word is a caption word)
                If the current character is the first character in the series numbering
                    Replace the current character with the caption substitute
                Else
                    Temporarily normalize the portion of the series numbering that
                            precedes the current character
```

[65] Example: in the numbering ‘1 (O.T. 1)’, the ‘T’ is retained.

[66] Example: in the numbering ‘A.N.M.S. 90A’, the ‘N’ is retained.

[67] Example: the numbering ‘T. XVIII, fasc. 6’ becomes ‘ | . XVIII, fasc. 6’

[68] Do *not* use the caption substitute.

[69] Example: in the numbering ‘T.P., no. 1’, retain the ‘P’.

[70] Do not use the caption substitute. Example: the numbering ‘3. F., Nr. 46’ becomes ‘3. SER ., Nr. 46’

```
                            If the last word in this normalized portion of the series numbering is
                                    not present in the list of multiple-character caption words
                                Replace the current character with the caption substitute
            Else (the current character is the lowercase form)
                Replace the current character with the caption substitute[71]
        Else (character following the current character is not one of those listed above)
            Retain the current character[72]
```

Series numbering	**Becomes**[73]
`t. 3.`	`| . 3`
`Bd. 4, lfg. N (T. 1)`	`Bd. 4, lfg. N (| . 1)`
`2e s´er., t. 51.`	`2e s´er., | . 51.`
`L.P. 56`	`L.P. 56`
`1 (O.T. 1)`	`1 (O.T. 1)`

Letters ‘v’ and ‘V’

Perform the work described here for the letter in its uppercase form, and separately for the letter in its lowercase form.

```
Scan the series numbering from right to left
If the current character is the character of interest
    If the current character is the first character in the series numbering[74]
        If the character of interest is ‘v’ (lowercase)
            If the character following the current character is a numeral
                Replace the current character with the caption substitute[75]
        If the character following the current character is a space
            If the character of interest is ‘v’ (lowercase)
                If the character following the space is a numeral
                    Replace the current character with the caption substitute[76]
                Else if the character following the space is a full stop
                    If the character following the full stop is a space and if the character following
                            this space is a numeral
                        Replace the current character with the caption substitute[77]
                    Else if the character following the full stop is a numeral
                        Replace the current character with the caption substitute[78]
            If the character following the space is not a numeral or a full stop
                Temporarily apply standard normalization to the portion of the series numbering
                        that follows the space, and extract from this the first word
                If this extracted word is in the list of multi-character caption words
                    Replace the current character with the caption substitute[79]
        Else if the character following the current character is a full stop
```

[71] For example, ‘t. I, 1’ becomes ‘ | . I, 1’; ‘d. nr. 12’ becomes ‘ | . nr. 12’

[72] For example, the ‘R’ is retained in ‘no. R0054’; the ‘N’ is retained in ‘Bd. 4, Lfg. N2’.

[73] These examples show the numbering only as normalized up to this point.

[74] This sequence of instructions will also be followed for most cases of the character occurring *within* the number, if certain conditions set forth below are met.

[75] Example: ‘v15’ becomes ‘ | 15’.

[76] Example: ‘v 5’ becomes ‘ | 5’.

[77] Example: ‘v . 3’ becomes ‘ | . 3’.

[78] Example: ‘v .3’ becomes ‘ | .3’.

[79] Example: ‘v no. (PHS) 80-3279’ becomes ‘ | no. (PHS) 80-3279’.

```
    If the character of interest is ‘V’ (uppercase)
      If none of the words in the series numbering appears to be a roman numeral[80]
        If the current character is not the first character in the series numbering and if the character preceding the ‘V’ is a full stop
          Retain the current character[81]
        If the character following the full stop is another uppercase letter and if the character following this second uppercase letter is not another full stop
          Replace the current character with the caption substitute[82]
        Temporarily normalize the portion of the series numbering that follows the ‘V.’ and isolate from this the first word
        If this isolated word is not in the list of multi-character caption words
          If the current character is not the first character in the series numbering
            Temporarily normalize the portion of the series numbering that precedes the ‘V.’ and isolate from this the last word
            If the word preceding the ‘V’ is not in the list of multi-character caption words
              If this previous word ends with a full stop itself preceded by a digit and if the following word begins with a full stop
                Retain the current character[83]
              Else
                Replace the current character with the caption substitute[84]
        Else (isolated word following ‘V.’ is in the list of multi-character caption words)
          If the current character is the first character in the series numbering
              Replace the current character with the caption substitute
          Else (current character is not the first character)
            Temporarily normalize the portion of the series numbering that precedes the ‘V.’ and isolate from this the last word
            If this isolated word is not in the list of multi-character caption words
              Replace the current character with the caption substitute
    Else (the character of interest is’v’—lowercase)
      Temporarily normalize the portion of series numbering that precedes the current character (do not isolate any word from this string)
      If this normalized numbering fragment is just ‘V’
        Retain the current character
      Else if the first character in this normalized numbering is ‘V’ and if the entire normalized fragment of the series numbering is in the list of multi-character caption words (which should only happen if this normalized fragment contains a single word)
        Retain the current character[85]
      Else
        Replace the current character with the caption substitute
Else if the current character is the last character in the series numbering
  Retain the current character[86]
```

[80] Apply this test to each word in a temporarily-normalized version of the series numbering: If *all* of the letters in any word are potential roman numeral letters (IVXLCDM), assume that the series numbering contains a roman numeral. (If a word in the series numbering is just ‘V’, do not include the word in this test. Recall that this test is being performed precisely because an ambiguous isolated character ‘V’ has been found in the series numbering.)

[81] Example: the ‘V’ in ‘1931.V.7’ is not changed.

[82] The corpus of series headings contains no example of this condition.

[83] Example: retain the ‘V’ in the numbering ‘1926. V. 13’.

[84] Example: the numbering ‘Ser. 2, V. 5’ becomes ‘Ser. 2, | . 5’.

[85] Example: the ‘v.’ in ‘vol. v. no.2’ is not changed.

```
Else (current character is within the series numbering)
    If the character preceding the ‘v’ or ‘V’ is a semicolon, full stop, comma, opening
            parenthesis, opening brace, opening bracket, less than sign or space
        If the character of interest is ‘V’ and occurs as the third or later character in the series
                number and is preceded by a space which is itself preceded by a digit, and is
                followed by a space plus another digit
            Retain the current character[87]
        Else
            Proceed as described above for ‘v’ or ‘V’ that occurs as the first character
    Else if the character preceding the ‘v’ is a hyphen
        Proceed as described above for ‘v’ or ‘V’ that occurs as the first character
    Else
        Retain the current character
```

3.5.2.6 Preliminary character handling

The instructions in this section apply to all instances of series numbering. This stage entails most but not all of what might properly be called ‘normalization’ (even though the normalization performed here is not the same as NACO normalization). The final part of normalization is described in sections 3.5.3.2 and 3.5.4.3.3.

```
Scan the series numbering from right to left
For each character:
    If the current character is a left bracket, left parenthesis, left brace, less than sign or equals
            sign
        If the current character is the first character in the series numbering
            Remove the current character
        Else
            Replace the current character with a semicolon and a space[88]
    Else if the current character is a right bracket, right parenthesis, right brace, greater than
            sign, colon, full stop, ‘pound’ sign, plus sign or ampersand
        Replace the current character with a space
    Else if the current character is a backward slash
        Replace the current character with the forward slash
    Else if the current character is the forward slash, comma, semicolon, space, uppercase letter
            of the alphabet or caption substitute
        Retain the current character
    Else if the current character is a hyphen
        If the hyphen is the first character in the series numbering
            Perform Normal punctuation handling
        Else if the hyphen is the second character in the series numbering
            Perform Internal hyphen handling
        Else if the hyphen is the third character in the series numbering
            If the first 2 characters in the series numbering are ‘al’
                Remove the first 3 characters in the series numbering[89]
            Else
```

[86] Example: the ‘V’ in ‘Bd. 3, Kapitel V’ is not changed.

[87] Example: the ‘V’ in ‘FM 10-76 V 3/4.’ is not changed

[88] Examples: ‘Bd. 2, [ERGANZUNGSBAND’ becomes ‘Bd. 2, ; ERGANZUNGSBAND’. At a later point in the normalization of the series numbering (section 3.5.4.1), this semicolon causes the portion of the numbering appearing to the left of this punctuation to be treated as separate from the portion appearing to the right.

[89] Examples: ‘al-ADAD 3’ becomes ‘ADAD 3’; ‘al-MUJALLAD X’ becomes ‘MUJALLAD X’.

```
                Perform Internal hyphen handling
        Else if the hyphen is the last character in the series numbering
            Perform Normal punctuation handling
        Else (the hyphen is somewhere else within the series numbering)
Internal hyphen handling:
            If the three characters preceding the hyphen are space plus ‘al’
                Remove the ‘al’ and hyphen (leaving the space)
            If the character preceding the hyphen is a digit
                If the next significant character following the hyphen is a letter of the alphabet
                    Perform Normal punctuation handling
            Else if the character preceding the hyphen is a letter of the alphabet
                If the character following the hyphen is not a numeral
                    If the character following the hyphen is not a letter of the alphabet
                        Replace the hyphen with a space[90]
                Else if the character following the hyphen is a letter of the alphabet
                    Perform Normal punctuation handling
            Else (the character preceding the hyphen is some other character)
                Perform Normal punctuation handling
    Else if the current character is a lowercase letter of the alphabet
        Replace the current character with its uppercase equivalent
    Else if the current character is a digit
        If the digit is not the first character in the numbering
            If the character preceding the digit is a superscript or subscript character (including punctuation)
                Insert a space before the current character, separating it from the superscript or subscript character[91]
    Else if the current character is a superscript or subscript character
        If the character is not the first character in the numbering
            If the character preceding the superscript or subscript character is a digit
                Insert a space before the current character, separating it from the digit, and replace the current character with its standard normalized equivalent[92]
            Else
                Replace the current character with its normalized equivalent
        Else
            Replace the current character with its normalized equivalent
    Else if the current character has an ASCII value of 128 or higher (diacritics and special characters)
        Replace the current character with its standard normalized equivalent
    Else (the current character is some mark of punctuation not mentioned above)
Normal punctuation handling:
        If the current character is the first or last character in the series numbering
            Remove the current character[93]
        Else
            If the character following the current character is a numeral
                If the first significant character preceding the current character is a numeral
                    Replace the current character with a space[94]
                Else
                    Remove the current character[95]
            Else
```

[90] Examples: ‘n.F., 1 a- B’ becomes: ‘n.F., 1 A B’; ‘P- 2128’ becomes ‘P 2128’

[91] Example: ‘sv. 245$_2$46’ becomes ‘sv. 245$_2$ 46’.

[92] Examples: ‘sv. 245$_2$ 46’ becomes ‘sv. 245 2 46’; ‘61$_2$, 146’ becomes ’61 2, 146’

[93] Example: ‘Bd. 33-‘ becomes ‘Bd. 33’.

[94] Example: ‘CR-88-01’ becomes ‘CR-88 01’.

[95] Example: ‘CD-75’ becomes ‘CD75’

```
            Remove the current character
```

The following examples show selected instances of series numbering as modified up to this point.

Numbering from subfield ǂv	Becomes
37	37
v. 54	| 54
reel 23, no. 2	REEL 23, NO 2
new series, v. 21	NEW SERIES, | 21
Tomus VI	TOMUS VI

3.5.3 Concluding steps for series numbering that contains no digits

3.5.3.1 Introduction

Series numbering that does not contain digits[96] cannot be handled in the same manner as series numbering that contains digits. It is much more difficult to identify those parts of the numbering that may constitute caption information when there are no reliable anchor points within the string. An attempt to identify caption words in these strings brings with it the likelihood that important characters will be omitted. Instead, the handling for series numbering that does not contain numerals aims principally at the reduction of information to a standard form without introducing much opportunity for damage.

3.5.3.2 Final character normalization

Technique:
Perform standard normalization on the series numbering.

3.5.3.3 Substitution for abbreviations, etc.

```
Technique:
For each word in the normalized series numbering
    If the word is contained in a list of words that means ‘new series’
        Replace it with the ‘new series’ replacement text
    Else if the word is contained in a list of words that means ‘new’ and is either preceded or
            followed in the segment by a word that means ‘series’
        Replace the ‘new’ and ‘series’ words with the ‘new series’ replacement text
    Else if the word is contained in a list of words that means ‘series’ and is preceded or followed
            in the segment by a word that means ‘new’
        Replace the ‘new’ and ‘series’ words with the ‘new series’ replacement text
    Else if the word is not part of a pair that means ‘new series,’ and if the word is found in the list
            of words that are to be reduced to a standard form
        Replace the word with its standard form
```

[96] The corpus of 696,510 series headings contains 2,914 headings (0.42%) whose numbering contains no digits.

Numbering	Becomes
ABDRUCK F	ABD F
NY SERIE I C	NEW SER I C

3.5.3.4 Examples

The following examples illustrate the complete normalization of series numbering that does not contain digits, as determined by all of the pertinent instructions in sections 3.5.2 and 3.5.3.

Numbering	Becomes
tomus IX	T IX
extra volume	EXTRA V
part C	PT C
levels B-F	LEVEL B F
level D	LEVEL D
VIII bd., n° III	VIII BD III

3.5.4 Concluding steps for series numbering that contains digits

3.5.4.1 Introduction

The intent of these instructions is to remove from the series numbering as much text that appears to be caption information as possible, without removing information that is not caption information. Text that remains in the numbering is regularized to the extent possible.

An instance of series numbering is not necessarily processed all at once, but is divided into *chunks*. An instance of series numbering may consist of a single chunk, or of any number of chunks. In most cases, each comma, semicolon or slash marks the dividing point between two chunks. (The punctuation itself is not part of any chunk; it becomes a space in the normalized form of the series numbering.) However, if the last chunk does not contain any digits, it is included as part of the preceding chunk.

Numbering[97]	Chunks
`BD 13, PT 3`	`BD 13` *and* `PT 3`
`N F , BD 27`	`N F` *and* `BD 27`
`2ND SER , 4, ETC`	`2ND SER` *and* `4 ETC`[98]
`11A; 11A, PT 1; 11A, PT 2AD`	`11A` *and* `11A` *and* `PT 1` *and* `11A` *and* `PT 2AD`
`36, NO 1/2 JAN /MARCH, 1972`	`36` *and* `NO 1` *and* `2 JAN`[99] *and* `MARCH` *and* `1972`

[97] The text in this column shows the series numbering as normalized up to this point.

[98] The final alphabetic chunk, ‘ETC’, is included with the previous chunk because the final chunk contains no digits.

BD 20	BD 20
NO DOT/RSPA/DPB/50/78/21-22	NO DOT *and* RSPA[100] *and* DPB *and* 50 *and* 78 *and* 21-22
78/98	78 *and* 98

First apply to the numbering the instructions in section 3.5.2. Then apply the instructions in sections 3.5.4.2, 3.5.4.3 and 3.5.4.4 (in that order) to each chunk in the series numbering. Instructions may apply to a chunk only when it is, or is not, the final chunk in the series numbering.

After applying these instructions to each chunk, reassemble the series numbering from the modified form of each chunk (supplying a blank space between each chunk), and then apply to the series numbering the instructions in section 3.5.4.5.

3.5.4.2 Conversion of ordinal numbers

The following scheme translates English-language ordinal numbers, and a few non-English ordinal numbers,[101] into their cardinal equivalents. Schemes may be developed along similar lines for ordinal numbers written in other languages if experimentation demonstrates that this transformation may safely be achieved without disordering series entries.[102]

```
For each word in the chunk of series numbering
    If the word is more than one character long, begins with a digit and contains a subsequent
            alphabetic character
        If this is the last word in the chunk and if the word contains no numerals and if the chunk
                is the last chunk in the series numbering
            The word represents a 'potential single letter problem'
        Temporarily divide the word at the first alphabetic character into numeric and alphabetic
                parts[103]
        If the last character in the numeric portion is '1'
            If the remaining portion of the word is 'ST' (as in '21st') or 'TH' (as in '11th')
                Replace the current word with just its numeric portion
            Else if the remaining portion of the word is 'RE' (as in '1re') or 'ERE'
```

[99] '2 JAN' is an example of a chunk that is bounded on both ends by slashes. This distinction is referred to at one point in the following instructions.

[100] RSPA, DPB, 50 and 78 are examples of chunks bounded on both ends by slashes.

[101] In the scheme described here, non-English ordinal numbers are only translated to their cardinal form if they are part of a numbered subseries designation or other recognizable construction likely to contain such a number.

[102] Examples of series numbering that contain information that appears to be an ordinal number, but isn't: 'v. 28a' (in this series, there is also a 'v. 28'); 'reel 22, no. 14a' (in this series there is also a 'reel 22, no. 14').

[103] Examples:

Word from series numbering	Temporarily treated as
2ND	2 *and* ND
15C	15 *and* C
3B5	3 *and* B5

```
        If the preceding or following word in the numbering is in the list of words
                associated with ordinal numbers
            Replace the word with just its numeric portion
    Else if the remaining portion of the word is ‘A’, ‘E’ or ‘O’ (as in ‘1a’, ‘2o’ or ‘1e’)
        If the word does not present a ‘potential single letter problem’
            If the preceding or following word in the numbering is in the list of words
                    associated with ordinal numbers
                Replace the current word with just its numeric portion
Else if the last character in the numeric portion is ‘2’
    If the remaining portion of the word is ‘ND’ (as in ‘32nd’) or ‘TH’ (as in ‘12th’)
        Replace the word with just its numeric portion
    Else if the remaining portion of the word is ‘ME’ (as in ‘2me’), ‘EME’ or ‘D’ (as in
            42d)[104]
        If the preceding or following word in the numbering is in the list of words
                associated with ordinal numbers
            Replace the current word with just its numeric portion
    Else if the remaining portion of the word is ‘A’, ‘E’ or ‘O’
        If the word does not present a ‘potential single letter problem’
            If the preceding or following word in the numbering is in the list of words
                    associated with ordinal numbers
                Replace the current word with just its numeric portion
Else if the last character in the numeric portion is ‘3’
    If the remaining portion is ‘RD’ (as in ‘103rd’) or ‘TH’ (as in ’13th’)
        Replace the word with just its numeric portion
    Else if the remaining portion of the word is ‘ME’ (as in ‘3me’), ‘EME’ or ‘D’ (as in ‘43d’)
        If the preceding or following word in the numbering is in the list of words
                associated with ordinal numbers
            Replace the current word with just its numeric portion
    Else if the remaining portion of the word is ‘A’, ‘E’ or ‘O’
        If the word does not present a ‘potential single letter problem’
            If the preceding or following word in the numbering is in the list of words
                    associated with ordinal numbers
                Replace the current word with just its numeric portion
Else if the last character in the numeric portion is ‘4’ through ‘0’
    If the remaining portion of the word is ‘TH’
        Replace the word with just its numeric portion
    Else if the remaining portion of the word is ‘ME’ (as in ‘5me’) or ‘EME’
        If the preceding or following word in the numbering is in the list of words
                associated with ordinal numbers
            Replace the current word with just its numeric portion
    Else if the remaining portion of the word is ‘A’, ‘E’ or ‘O’
        If the word does not present a ‘potential single letter problem’
            If the preceding or following word in the numbering is in the list of words
                    associated with ordinal numbers
                Replace the current word with just its numeric portion
```

Chunk as received	Chunk with ordinal number converted
3RD SER	3 SER
26TH	26
SER 1A	SER 1
93RD CONGRESS	93 CONGRESS

[104] *Rules for descriptive cataloging*, 1949 Appendix IV.G and AACR1 Appendix IV.H mandated the use of the abbreviations ‘2d’ and ‘3d’ for the ordinal numbers ‘second’ and ‘third’.

Make no attempt to convert the spelled-out forms of ordinal numbers in any language ('first', 'deuxième', 'dritte') into digits unless careful analysis is undertaken of the surrounding text.[105] These forms occur rarely enough that such extensive analysis is not likely to be repaid by greatly improved ordering of series headings.

3.5.4.3 Segmentation

3.5.4.3.1 Introduction

After being inspected for ordinal numbers, the series numbering chunk is further divided into *segments*. Each transition from a non-numeric character (alphabetic character, mark of punctuation or space) to a digit, or from a digit to a non-digit, marks the dividing line between segments. Spaces that occur between digits also mark the dividing point between segments, but spaces that occur between non-digits do not divide segments. Separate handling is defined for segments composed of digits and segments that contain no digits. Introduce a space (if not already present) at each segment boundary within the chunk.

If the dividing point between the current and previous segment is not a space and if the segment to the left of the dividing point is numeric (which means that the segment to the right must be non-numeric), temporarily label the first word in the following, alphabetic segment as requiring special treatment.[106] Similarly, if the dividing point between the current and previous segment is not a space and if the segment to the left of the dividing point is not numeric (which means that the next segment must be numeric), temporarily mark the final word in the alphabetic segment as requiring special treatment.

The following examples show the division of chunks of series numbering into segments.

Series numbering in subfield ǂv	**Becomes these segments**
Bd. 1	BD *and* 1
no. 24744-24747	NO *and* 24744 *and* 24747
91-14S	91 *and* 14 *and* -S[107]
DS5-S1	DS- *and* 5 *and* -S-[108] *and* 1

[105] Consider this example: 'of the twenty-second series the one-hundred-and-thirty-seventh volume'.

[106] The program used to test this algorithm added a hyphen to the indicated word in the alphabetic segment to mark the need for special treatment. This hyphen serves as a *blocking character* that prevents the word to which it is attached from matching any of the words in the lists of words to be omitted or modified. (This blocking character is eventually removed from the word.) Any mark of punctuation that has by this point been removed from the heading could be a candidate for use as the blocking character; any gimmick that achieves the same end as the blocking character (preventing a given word from being found in a list of words) could be substituted. For example, the word could be converted from uppercase to lowercase, and later converted back into uppercase. The examples in this report show the use of the hyphen as the blocking character.

[107] The hyphen shown at the beginning of this word represents the blocking character described in this section.

[108] The first hyphen (present in the original series numbering) is retained with the associated alphabetic segment. The second hyphen is the blocking character, added because the following numeric segment was

1981 no. 10	1981 *and* NO *and* 10
DP 97-002	DP *and* 97 *and* 002
9A supplementum	9 *and* -A SUPPLEMENTUM
ATLA fiche 1985-0350	ATLA FICHE *and* 1985 *and* 0350

3.5.4.3.2 Handling of numeric segments

Follow the instructions in section Appendix C to modify each numeric segment in a manner that allows it to be sorted as a value instead of as a set of characters.

3.5.4.3.3 Handling of alphabetic segments

```
If this alphabetic segment comes from the last chunk in the series numbering and the chunk was not divided into segments or if the alphabetic segment constitutes an entire chunk from the series numbering that was originally bounded at both ends by slashes
    Perform system normalization on the segment
    Perform the substitutions described below for 'new series' and standardized forms of caption words
    Perform none of the other work described in this section on this segment

If the segment consists solely of one of the 'series' words[109] and contains 4 or more characters, take one of the following actions
    If the portion of the series numbering to the left of this segment contains no space
        Add the 'series' replacement word to the beginning of the series numbering, and remove the current segment
    Else if the last word in the series numbering begins with a numeral
        Add the 'series' replacement word just before this numeral, and remove the current segment
    Else
        Replace the current segment with the 'series' replacement word
    Perform none of the other work described in this section on this segment

Perform the following tests on each word in the segment, from left to right. Do not perform this work if excluded by preceding instructions in this section
    If the word is the caption substitute
        Remove the caption substitute and set the 'word removed' flag[110]
    If the word is in the list of 'new' words
        The list of 'other' words is the list of series words
        The 'replacement text' is the 'new series' replacement
        Perform Handle series word
    Else if the current word is in the list of 'hors' words
        Set the 'hors' flag
```

directly attached to this segment; even were the first hyphen not present, this second hyphen would prevent this word from matching any word in any of the lists of words. All punctuation will eventually disappear.

[109] Here and elsewhere, recall that, during the division of a chunk into segments, when an alphabetic character is immediately adjacent to a numeric character, the alphabetic character is marked in some way that prevents the word that contains it from meeting tests for inclusion in various lists of words; the word will be retained in the series numbering. (For example, if the series numbering in subfield ╪v is NR171, the segment 'NR' will not be found in any list of words of interest—even though 'NR' is a caption word—and will be retained in the series numbering.)

[110] At a later point in this section, the fact that the caption substitute has just been removed will (as will the removal of a multi-character caption) become important.

```
        The list of ‘other’ words is the list of series words
        The ‘replacement text’ is the ‘hors series’ replacement
        Perform *Handle series word*
    Else if the current word is in the list of ‘new series’ words
        The ‘replacement text’ is the ‘new series’ replacement
        Replace the current word with the ‘new series’ replacement
        Clear the ‘word removed’ flag
        Perform *Handle series word*
    Else if the current word is in the list of ‘series’ words
        Set the ‘series’ flag
        The list of ‘other’ words is the list of ‘new’ words
        The ‘replacement text’ is the ‘hors series’ replacement
        If the first four letters of the final word in the series numbering as normalized up to this
                point are in the list of ‘whole’ words
            Retain the ‘series’ word in the heading
            Clear the ‘word removed’ flag
        If the last character of the series numbering as normalized up to this point is not a digit
            If the first character in the remainder of the series numbering is not a digit *or* if the
                    ‘previous word was removed’ flag is set
                Retain the current word in the numbering
                Clear the ‘word removed’ flag
        Perform *Handle series word*
    Else if the current word is in the list of multi-character caption words
        If the word is in the list of words that constitute the first part of an ordinal numeral in
                Chinese, Japanese or Korean series numbering
            Set the ‘CJK exception’ flag
            Remove the current word from the series numbering
            Set the ‘word removed’ flag
    Else
        If the word is in the list of words that constitute the first part of an ordinal numeral in
                Chinese, Japanese or Korean series numbering[111]
            Set the ‘CJK exception’ flag
            Remove the current word from the series numbering
            Set the ‘word removed’ flag
            Continue with the next word in the segment
        If the ‘CJK exception’ flag is set[112]
            Clear the ‘CJK exception’ flag
            If the current word is in the list of words that constitute the second part of an ordinal
                    numeral in Chinese, Japanese or Korean series numbering[113]
                Remove the current word from the series numbering
                Set the ‘word removed’ flag
```

[111] This may at first seem redundant with a preceding test, but as it happens not all of the CJK ordinal number words can also be included in the list of caption words. Specifically, because the word ‘DI’ also appears in Italian-language captions (and should be retained), this second test just for CJK words must be performed here.

[112] Note that this does not simply refer to a previous word in the segment or chunk, but to any previous word in the series numbering as normalized up to this point. The indication that a CJK ordinal label was removed is reset whenever an alphabetic word is *not* removed, but is not reset due to the presence of intervening numerals.

[113] The following examples of series numbering illustrate the two-part ordinal numbers that are handled correctly by the technique described in this section:

```
dai 10-kan
ti 2 chi
di 16 zhong
```

```
    If the word is in the list of words for which substitutions are provided *and* if the chunk was not delimited in the original series numbering by a slash at either end
        Replace the word with its substitute and clear the ‘word removed’ flag
    Else
        Perform standard normalization on the current word
        Retain the current word in the numbering (as normalized) and clear the ‘word removed’ flag
```

Handle series word:

```
If this is the last word in the segment
    If the word is one character long
        Retain the word in all cases[114]
    Else
        If the word is not contained in the list of ‘other’ words and contains more than one character
            If the ‘series’ flag is set
                Set the ‘series removed’ flag
        Remove the word from the segment
Else (not the last word in the segment)
    If the next word in the segment is in the list of ‘other’ words
        Replace the current word and the next word with the replacement text
        Continue with the next word in the segment
    If the ‘series’ flag is set
        Clear the ‘series’ flag
        If the last word in the series numbering as normalized up to this point is in the list of ‘hors’ words
            Replace the final word in the series numbering as normalized up to this point with the ‘hors series’ replacement
        If the current word is one character long
            Retain the word in the series numbering
        Else
            Set the ‘series removed’ flag
            Do not retain the word in the series numbering
        Continue with the next word in the segment
    Else if the ‘hors’ flag is set
        Clear the ‘hors’ flag
        If the final word in the series numbering as normalized up to this point is in the list of ‘hors’ words
            Replace the current word and the final word in the series numbering as normalized up to this point with the ‘hors series’ replacement
    Else
        Normalize the current word
        Retain the current word in the segment
```

The following examples show the handling of alphabetic segments within series numbering.

Series numbering	**Becomes**
`81H`	`00000081 H`

[114] One-character words that need to be removed were handled at an earlier point. Any remaining one-character words should be preserved.

83E	00000083 E
AC 150/5390-1V	AC 00000150 00005390 00000001 V

The fact that terminal alphabetic segments are handled differently from other alphabetic segments means that the same word appearing in different segments within one instance of series numbering may be handled differently.

Numbering	Becomes:
80-38 (part I), 80-38 (part II)	00000080 00000038 I 00000080 00000038 PT II

3.5.4.4 Re-insertion of ‘series’ abbreviation

During the processing of a chunk of series numbering, a word that means ‘series’ may have been removed. The following procedure inserts a replacement for the deleted word at a standard location within the chunk. This allows variant forms of headings for numbered series to be reconciled.

```
If during the handling of a chunk the ‘series removed’ flag was set
    Clear the ‘series removed’ flag
    If the ‘series’ abbreviation is not otherwise present in the portion of the series numbering
            normalized from the chunk
        If this is the first chunk in the series numbering
            Add the ‘series’ abbreviation to the beginning of the series numbering
        Else if any other characters from the current chunk were retained in the series numbering
            Insert the ‘series’ abbreviation into the series numbering at the start of the current
                    chunk
        Else
            Insert the ‘series’ abbreviation into the series numbering at the point formerly
                    occupied by the current chunk

If the normalized chunk ends with a space plus the series replacement
    If the series abbreviation is preceded by the word ‘A’, ‘E’ or ‘0’[115] which is itself preceded by a
            word that begins with a numeral
        Insert the series abbreviation before the numeric word, and remove the ‘A’, ‘E’ or ‘0’
    Else if the word to the left of the series replacement begins with a digit
        Insert the series replacement before of the numeric word[116]
```

Chunk from series numbering	Becomes
ser. 1a	SER 00000001
2. Reihe	SER 00000002
2a sér.	SER 00000002
3rd ser.	SER 00000003

[115] The test for ‘A’, ‘E’ and ‘O’ is intended to catch ordinal numbers including numbered subseries not caught in previous tests.

[116] This test is necessary because text following the last digit is not manipulated.

3.5.4.5 Removal of abbreviation for the series heading

If the series numbering as prepared according to the preceding instructions still begins with any character other than a digit, subject the alphabetic prefix to the numbering (the portion that precedes the first digit) to additional tests. The aim of these tests is to remove from the numbering recognizable representations (typically, initials) of the series title. This representation is sometimes included by the publisher and recorded in the series statement by the cataloger, and sometimes it is not so included or recorded; removing it whenever possible allows additional series to be sorted correctly.

```
440  0 ǂa Twayne’s world authors series ; ǂv TWAS 19
440  0 ǂa Twayne’s world authors series ; ǂv 23
```

Do not apply these manipulations to series numbering that begins with the words ‘NEW’, ‘HORS’, ‘SUP’ or ‘SER’; to numbering whose alphabetic prefix consists of a single letter; or to numbering that begins with a digit.

The ‘manipulated version of the series numbering prefix’ referred to in this section is the alphabetical portion of the normalized series numbering, up to but not including the first digit, with all spaces (including internal spaces) removed.

Extract several sets of initials from the normalized form of the series heading:

- one set from the first letter of each word in the normalized heading
- one set from the first letter of each word in the normalized heading other than words in the list of ‘short’ words[117]
- if the non-normalized form of the heading contains a hyphen, one set of initials from the first letter of each word in the non-normalized heading (counting the hyphen as the beginning of a word)
- if the non-normalized form of the heading contains a hyphen, one set of initials from the first letter of each word in the non-normalized heading (counting the hyphen as the beginning of a word) other than words in a list of ‘short’ words

Series heading	**Extracted initials**[118]
`Biblioteca de autores cristianos`	`BDAC BAC`
`Harvest book`	`HB`
`Excavations at Dura-Europos`	`EADE EDE`

```
If the manipulated version of the series numbering prefix is not ‘SER’ and is more than 1
        character long and is contained anywhere within any of the four sets of initials
    Remove the alphabetic prefix from the series numbering
```

Series heading	**Normalized numbering**
`Biblioteca de autores cristianos ; ǂv BAC 65`	`00000065`
`Youth forum series ; ǂv HF 20`	`00000020`
`Harvest book ; ǂv HB277`	`00000277`

[117] These first two sets of initials are referred to again at the very end of this section.
[118] In all of the examples in this section, redundant sets of extracted initials are not shown.

```
If the non-normalized series heading contains an opening parenthesis
    Construct a new heading consisting of the original heading starting at the opening
        parenthesis, a space, and the original heading from the beginning to the opening
        parenthesis
    Normalize this new version of the heading
    Extract the four sets of ‘initials’ described above from this heading[119]
```

Series heading	Extracted initials
Technical report (University of Ibadan. Dept. of Agricultural Economics)	UOIDOAETR UIDAETR
South Asia series (East Lansing, Mich.)	ELMSAS
Report (Malawi. Water Resources Branch)	MWRBR

```
    If the manipulated version of the series numbering prefix is not ‘SER’ and is more than 1
        character long and is contained anywhere within any of the four sets of initials
        Remove the entire alphabetic prefix from the series numbering
```

Series heading	Normalized numbering
Technical memorandum (Johns Hopkins University. Operations Research Office) ; ǂv ORO-T-261	00000261
Working paper (Pan African Institute for Development. West Africa) ; ǂv WA/WP-82/4	00000082 00000004
Dissertation (Rand Graduate School) ; ǂv RGSD-146	00000146

```
If the entire series numbering prefix (with internal spaces intact) matches the series heading
    Remove the series numbering prefix
```

Series heading	Normalized numbering
NSRDS-NBS ; ǂv NSRDS-NBS 57	00000057

```
If the first word in the series numbering overlaps the beginning of the first word of the normalized
        series heading
    Remove the first word from the series numbering
```

Series heading	Normalized numbering
Capricorn books ; ǂv CAP 28	00000028

[119] A similar operation might be contemplated for series name/title headings: initials could be derived from the title plus the name. The corpus of series headings contains no case in which such an operation resulted in a change to the series numbering.

```
Else if the first word in the series numbering overlaps the beginning of any other word of the
        series heading *and* the first word does not consist solely of characters that comprise
        roman numerals[120]
    Remove the first word from the series numbering
```

Series heading	Normalized numbering
United States. ǂb Office of Justice Programs. ǂt OJP guideline manual ; ǂv OJP M 7100.1C	M 00007100 00000001 C
Publication SP ; ǂv SP-24	00000024
Wisconsin briefs from the Legislative Reference Bureau ; ǂv brief 95-5	00000095 00000005

```
Extract the first word from the normalized form of the series heading, plus the initials of the
        remaining words
If the manipulated version of the series numbering prefix matches the beginning of these initials
    Remove the entire alphabetic prefix from the series numbering
```

Series heading	Normalized numbering
FJC staff paper ; ǂv FJC-SP-77-3	00000077 00000003
ONR Tokyo scientific monograph series ; ǂv ONRT M2	M 00000002
HP Laboratories technical report ; ǂvHPL-93-38	00000093 00000038

```
Extract the initials of the normalized form of the series heading, omitting initials of words that
appear to be articles
```

Series heading	Extracted initials
Department of the Army ROTC manual	DOARM
Basic concepts in the law of evidence	BCILOE

```
    If the manipulated version of the series numbering prefix is not 'SER' and is longer than 1
    character and matches any part of these initials
        Remove the entire alphabetic prefix from the series numbering
```

Series heading	Normalized numbering
Topics in the neurosciences ; ǂv TIN 6	00000006
State of the environment report ; ǂv SOE report no. 90-1	00004494

```
If the first word in the normalized series numbering is not composed entirely of letters used in
        roman numerals
    If the original version of the normalized series numbering alphabetic prefix contained any
            spaces[121]
        Do this in turn for each word in the remainder of the normalized series numbering prefix
```

[120] Numbering such as 'VIII.2.1929' is excluded by this test.

[121] Whether or not any of these words has been removed by instructions in this section.

```
If the word is composed entirely of letters used in roman numerals
    Discontinue all work on the series numbering prefix
If the word is one character long, or is 'SER', 'NEW' or 'HORS'
    Retain this word and the following part of the series numbering
    Discontinue further work on the series numbering prefix
If the word matches the beginning of the normalized heading
    Do not include the word in the final series numbering
    Continue with the next word
```

Series heading	Normalized numbering
Translation series (U.S. Atomic Energy Commission) ; ǂv AEC-tr-4494	00004494[122]

```
Else if the word is contained in neither of the first two sets of initials initially extracted
        from the heading
    Retain this word and the following part of the series numbering
    Discontinue further work on the series numbering prefix
```

Series heading	Normalized numbering
DHEW publication ; ǂv no. (ADM) 79-749	ADM 00000079 00000749[123]
Harper torchbooks ; ǂv TB13	TB 00000013

```
Else (the word is found in one of the sets of initials)
    Do not include the word in the final series numbering
    Continue with the next word
```

Series heading	Normalized numbering
NCAR technical note ; ǂv NCAR/TN-129+PROC	00000129 PROC
NASA conference publication ; ǂv NASA CP-002	00000002
Technical report (Air Force Flight Dynamics Laboratory (U.S.) ; ǂv AFFDL-TR-80-3019	00000080 00003019

Here are some additional examples illustrating the work described in this section.

Original series heading	Numbering normalized as
Cahiers de Cit´e libre ; ǂv CL-4	00000004
Dissertation (Rand Graduate School) ; ǂv RGSD-103	00000103
Report (University of Illinois at Urbana-Champaign. Dept. of Computer Science) ; ǂv no. UIUCDCS-R-86-1316	00000086 00001316

[122] The 'tr' is removed as a result of this instruction, 'AEC' having been removed in an earlier step

[123] 'no.' having been removed in an earlier step.

Information memorandum (United States. Administration for Children, Youth and Families) ; ǂv ACYF-IM-84-14	00000084 00000014
Air Force TO ; ǂv TO 13C7-1-5	00000013 C 00000007 00000001
ATLA series preservation program ; ǂv ATLA film 1996-S054	00001996 S 00000054[124]
Rehab/education resourcebook series ; ǂv resourcebook 4	00000004
Technical report (Old Dominion University. Dept of Geophysical Sciences) ; ǂv GSTR-84-12	00000084 00000012

3.5.4.6 Examples

The following examples show the handling of complete series numbering subfield ǂv from series headings when the subfield contains digits.

Series numbering in subfield ǂv:	**Becomes:**
Bd. 1	00000001
no. 24744-24747.	00024744 00024747
fiche 4,293-4,296	00000004 00000293 00000004 00000296
91-14S	00000091 00000014 S
30 ETC	00000030 ETC
1981 no. 10	00001981 00000010
DP 97-002	DP 00000097 00000002
Hft. 5. [1939]	00000005 00001939
N-1250-ARPA	N 00001250 ARPA
1, 2, plates	00000001 0000002 PLATES
80-R-4	00000080 R 00000004
33:2	00000033 00000002
87/3	00000087 00000003
Jan. 1983	JAN 00001983
n.F., Heft 55A-55B	NEW SER 00000055 A 00000055 B
n.F., Heft 38 A-B	NEW SER 00000038 AB
#9012	00009012
Heft 29	00000029
30 Heft	00000030 HFT
Jahrg. 1986, 3. Abhandlung, etc.	00001986 00000003 ABH ETC

[124] The word ‘film’ was removed from the series numbering in an earlier step because it appears in the list of caption words.

research report no. 3	00000003
96th Congress, 2nd session, no. 1	00000096 CONG 00000002 00000001
no. 3, 1st ser	00000003 00000001 SER
4th ser., 16	SER 00000004 00000016
68th Congress, no. 350	00000068 CONG 00000350
n.F., 2. Reihe, Bd. 27	NEW SER 00000002 00000027
Serie I, no. 2	SER I 000002
3. Folge, Nr. 143, etc.	00000003 00000143 ETC
80-38 (part I), 80-38 (part II)	00000080 00000038 I 00000080 00000038 PART II

4 Rating the normalization algorithms

4.1 Introduction

There are three principal measures of interest in the evaluation of an algorithm for normalizing series numbering:

1. How correct is the order of headings produced by the algorithm?
2. How much time does the algorithm require to normalize a heading?
3. How does the algorithm affect the size of the normalized heading?

The task group collected information to help answer all of these questions. The task group has no similar data related to a fourth measure of interest, namely the difficulty of designing, writing and testing programs that implement the algorithms. However, it is possible at least to say that techniques 1, 2 and 3 would be fairly easy to implement, and the implementation of technique 4 would be a major undertaking.

4.2 Standards for comparison

The results produced by the proposed normalization algorithms were compared among themselves, and also to two ‘standards:’ no normalization at all, and typical system normalization. These two schemes provide the baseline measurements against which the other routines can be evaluated.

‘Null’ normalization

‘Null’ normalization entails the use by the library system of the series numbering in subfield ╪v as found, without making any change of any kind. The test program for ‘null’ normalization entailed all of the overhead required of the other normalization algorithms, but made no change to the numbering itself.[125]

[125] The inclusion in the ‘null’ program of the overhead required of the other programs means that the differences between the timings for other techniques and the timing for the ‘null’ technique show the

Series numbering in subfield ǂv	Becomes
v. 5	v. 5
1997/1998	1997/1998
No. 53, 3d series	No. 53, 3d series
no. (PHS) 79-3216	no. (PHS) 79-3216

Naturally, this kind of 'normalization' produces no improvement in the order of series headings. All of the problems related to series numbering noted elsewhere are present in series headings arranged by the null algorithm.

System normalization

The library system performs the same normalization on series subfield ǂv that it performs on all other subfields in heading fields, and performs this normalization in exactly the same manner it does for those other subfields. The accompanying table shows selected series numbering from a few different headings as normalized by this technique.

Series numbering in subfield ǂv	Becomes
v. 5	V 5
1997/1998	1997 1998
No. 53, 3d series	NO 53 3D SERIES
no. (PHS) 79-3216	NO PHS 79 3216

This technique introduces no new mislocations, when compared to null normalization.

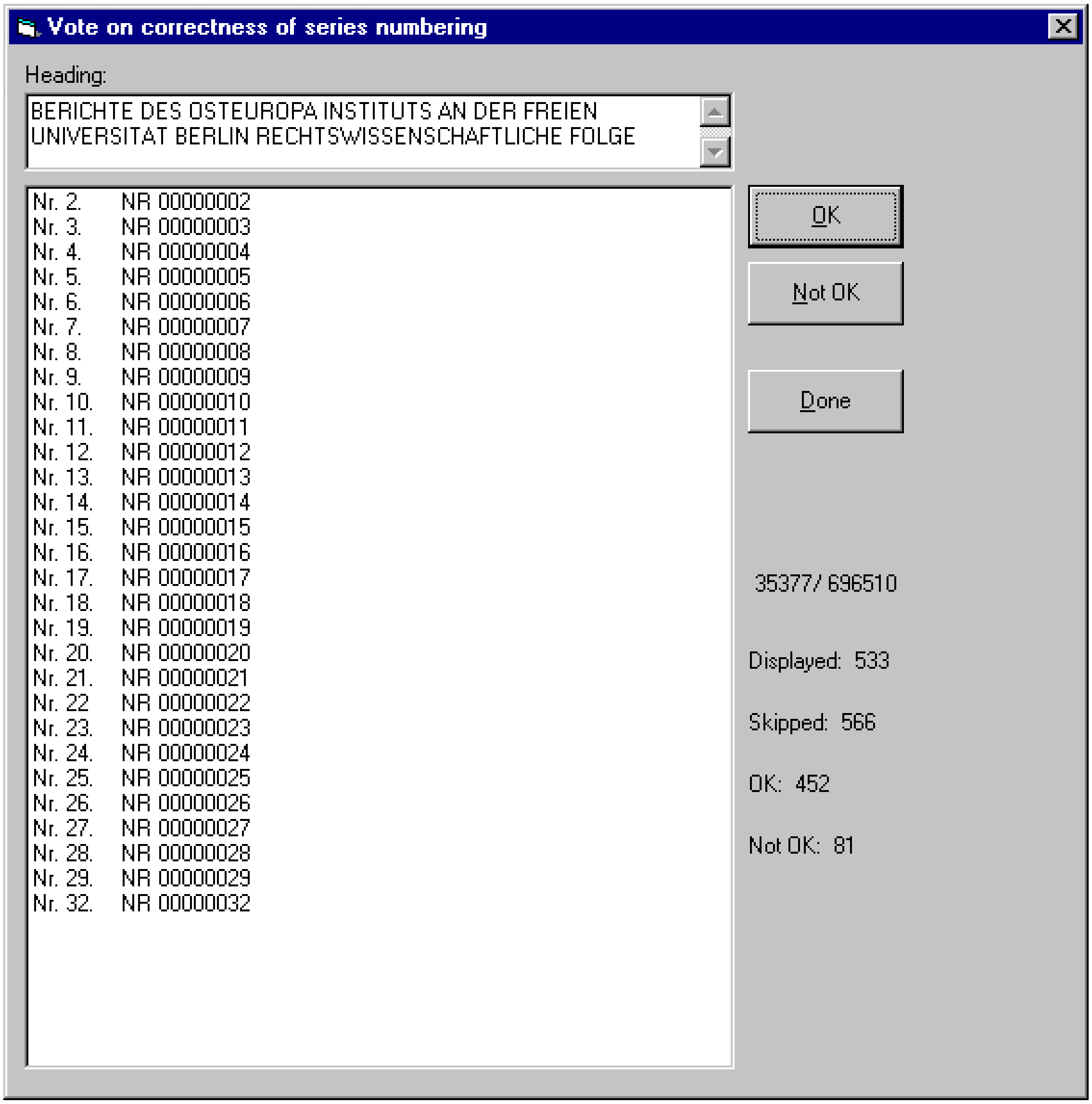
4.3 Correctness of results

The task group used a purpose-written program to help it judge the correctness of the order of headings produced by each proposed algorithm. This program takes as input the corpus of series headings, with the numbering normalized by some algorithm. The operator indicates how large the sample of headings should be;[126] the program selects headings from the corpus and presents a list of the members of each selected series for evaluation. The operator judges whether the entries for each heading are in the correct order or not.

amount of additional (non-overhead) time required by the other techniques—the amount of time required to do the work of the algorithm.

[126] To generate the ratings of series described this report, the requested sample contained 6550 headings. This is very nearly 8% of the headings in the corpus that occur more than once. (There is no need to evaluate the result produced by an algorithm on a given series if that series is represented by only one occurrence; the single heading is already 'in order.' Series headings that occur only once were processed by the normalization program and were often examined during the development of the algorithms, but were dropped from the tests for correct sorting.) Because the samples for all tests of the normalization techniques were the same size, the review program actually selected the *same headings* for each test.

The accompanying illustration shows the review program in operation. The first column in the large box shows the raw form of the series numbering, the second the normalized form. (The result set being examined was produced by the algorithm described in section 3.2—numerals are extended to a fixed length; non-numeric characters are normalized in the standard manner but are not otherwise manipulated.) The operator is expected to judge the correctness of the order of the headings, and to press either the 'OK' or 'Not OK' button. (For the series shown in this illustration, the operator should press the 'OK' button, because the members are properly arranged in ascending order.)



Heading: BERICHTE DES OSTEUROPA INSTITUTS AN DER FREIEN UNIVERSITAT BERLIN RECHTSWISSENSCHAFTLICHE FOLGE

```
Nr. 2.    NR 00000002
Nr. 3.    NR 00000003
Nr. 4.    NR 00000004
Nr. 5.    NR 00000005
Nr. 6.    NR 00000006
Nr. 7.    NR 00000007
Nr. 8.    NR 00000008
Nr. 9.    NR 00000009
Nr. 10.   NR 00000010
Nr. 11.   NR 00000011
Nr. 12.   NR 00000012
Nr. 13.   NR 00000013
Nr. 14.   NR 00000014
Nr. 15.   NR 00000015
Nr. 16.   NR 00000016
Nr. 17.   NR 00000017
Nr. 18.   NR 00000018
Nr. 19.   NR 00000019
Nr. 20.   NR 00000020
Nr. 21.   NR 00000021
Nr. 22    NR 00000022
Nr. 23.   NR 00000023
Nr. 24.   NR 00000024
Nr. 25.   NR 00000025
Nr. 26.   NR 00000026
Nr. 27.   NR 00000027
Nr. 28.   NR 00000028
Nr. 29.   NR 00000029
Nr. 32.   NR 00000032
```

35377/ 696510

Displayed: 533

Skipped: 566

OK: 452

Not OK: 81

The task group discovered that it was not always easy to determine whether the order of headings created as the result of the application of a given algorithm should be declared correct or incorrect; in some cases, a degree of judgment was called for. The task group recognized that absolute consistency in evaluating the order of headings was probably not an achievable goal, but it developed a few principles to help achieve a reasonable level of consistency.[127]

- Are the members of the series amenable to being into any 'right' order at all?[128] If no members of the series are amenable to being put into right order by algorithm, the members of the series are to be considered in order. If some members of the series are amenable to being put into order and some are not, judgment is passed only on those that are amenable to being put into order. Members amenable to being put into order should appear adjacent to each other, and should be in ascending order.
- Is more than one numbering pattern present? Entries that follow any one numbering pattern should appear adjacent to each other, and should be in ascending order.

The task group included in its category of correctly arranged headings those headings that were arranged correctly 'for the wrong reason.' Such headings occur under all of the algorithms described in this report. This category contains headings that are in the proper order by accident, despite the operation of the normalization algorithm.

Series numbering	Normalized form (section 3.2)
`75`	`00000075`
`núm. 110`	`NUM 00000110`

The headings are in the right order not because 75 comes before 110, but because '0' comes before 'N'; regardless of the reason, they are in the right order.

Series numbering	Normalized form (section 3.3)
`Ser. A, no. 6`	`00000006`
`Serie B, no. 28`	`00000028`

The headings are in the right order not because 'A' comes before 'B', but because '0' comes before '2'; regardless of the reason, they are in the right order.[129]

Similarly, the task group did not count as errors those series headings arranged correctly only because the corpus happened not to contain an entry whose presence would otherwise produce an error. Ratings were based only on headings in the corpus, not other headings that might possibly exist.

Series numbering	Normalized form (section 3.3)
`2e sér., t. 12`	`00000002 E SER T 00000012`
`2e sér., t. 51`	`00000002 E SER T 00000051`

[127] In an attempt to be as consistent as possible, all of the ratings of algorithms included in this report were made by the same person.

[128] Series numbering that contains roman numerals, numbers spelled as words, months and seasons of the year, and so on, are not amenable to being put into a right order by any algorithm. In some cases, one portion of an instance of series numbering is amenable to being put into order, and the remainder is not.

[129] It is perhaps of interest to point out that these two entries (the only entries in the corpus for this series) are placed in the correct order by all four of the normalization techniques discussed in this report. Levels 1 and 2 place them in the correct order by accident, levels 3 and 4 place them in the correct order by design.

> *These are the only two numberings present in the corpus for the series. If the corpus were also to contain, for example, 't. 5', it would, according to the normalization scheme being tested, incorrectly appear after the two entries shown above (its normalized form would be '00000005'). Because the corpus does not happen to contain such a numbering, the order of entries is correct as it stands.*

The accompanying table shows the correctness rating derived for each of the normalization algorithms described in this report.

Normalization technique	Percent correct	Increase over null routine
Null routine	53.91%[130]	—
System normalization	56.99%	5.73%
1	86.31%	60.13%
2	98.11%	81.99%
3	98.53%	82.80%
4	99.85%	85.25%

The application of system normalization does not produce a marked improvement in the number of correctly sorted headings. (The improvement is due chiefly to the general effects produced by normalization: the use of a single case for all alphabetic characters, the removal of punctuation, and the regularization of spacing.) Expanding numerals to a standard length (technique 1) brings the single greatest improvement to the arrangement of series headings. Disregarding text that appears to represent captions (technique 3) produces very good results, although surprisingly not much better than does disregarding text without testing it at all (technique 2). The most complex of the normalization techniques described in this report produces the best results; quite a bit of additional work is required to squeeze out this final improvement in the order of series headings from the remaining difficult cases.

The descriptions of the normalization algorithms in section 3 characterize the major areas each fails to address, and the new mislocations each creates. Although it preserves every mislocation caused by variations in the raw data, technique 1 creates no new mislocations, and is the only technique of the four that can make this claim. Techniques 2 and 3, when attempting to remove variations in captions, create mislocations by removing important text at the same time. Of the techniques that attempt to resolve variations in captions, technique 4 creates the smallest number of new mislocations.

[130] The program for examining series headings uses a database as its storage medium. Midway through the examination of series entries handled by the null routine, it was discovered that the database program was treating uppercase and lowercase forms of letters as the same value, instead of as distinct characters. If the test were to be re-performed with a database that did not treat lowercase and uppercase letters as the same value, the percent correct for the null routine would be slightly lower. This oversight does not affect the percentage correct found for the remaining algorithms, as they all involve the conversion of lowercase characters to their uppercase form. However, because the base percentage should be lower, the percentage *increases* for the remaining algorithms should be correspondingly higher.

4.4 Execution time

The programs that generated the normalized forms of series headings recorded the amount of time needed to process all of the headings in the corpus. These timings, summarized in the accompanying table, should not be taken as a guarantee of performance under a given algorithm, but only as relative guides to the amount of processing time consumed by programs implementing each algorithm.[131] Timings achieved in specific environments when using an implementation of a specific algorithm will of course vary, depending not only on system hardware but also on the degree to which the normalization routine has been optimized for speed of execution.

Normalization technique	Time needed to process all headings in the corpus[132]	Increase over null routine
Null routine	391 seconds	—
System normalization	451 seconds	15.35%
1	522 seconds	33.50%
2	532 seconds	36.06%
3	547 seconds	39.90%
4	857 seconds[133]	119.18%

The expansion of numerals to standard length (techniques 1, 2 and 3) consumes roughly twice as much excess time over null normalization as does system normalization. Although these three techniques vary somewhat in their timings (which are neatly sorted by the complexity of the work performed) they nonetheless form a cluster whose degree of spread should not be accorded undue weight, especially when it is remembered that the small difference in timing must be spread over nearly 700,000 headings. As might be expected, the most complicated algorithm (technique 4) required the most time to complete its task.

4.5 Size of normalized heading

The accompanying table summarizes the effect of each of the normalization algorithms on the length of the series numbering in the corpus. For the purposes of this table, the corpus contained 696,474 valid numbering subfields. The sizes of each of the valid numbering subfields created by each of the algorithms were added together to determine

[131] The normalization routines used for testing by the task group grew by accretion, and were not written from a well-considered *a priori* design. No claim is made that the code used to produce each set of normalized headings is the most efficient possible.

[132] Note for all of the timings given in this report that variations of a few seconds were commonly observed when performing the same test more than once. The timings given are the best observed for each algorithm. Algorithms that vary from each other in their timings by only a few seconds should be considered to have identical timings. Divided by the number of headings in the corpus (696,510), a difference of a few seconds becomes very small indeed. It may be of interest to note that skipping the work described in section 3.5.4.5 reduced the time required to process the corpus by only 20 seconds.

[133] An earlier, less efficient implementation of the technique 4 algorithm took 1564 seconds—nearly twice as much time—to process the headings in the corpus. As is the case with all of the routines in the test program, additional improvements in running time might be achieved through additional optimization.

the total number of characters occupied by the normalized numbering. Dividing this aggregate by the number of headings produces the average length of each numbering.

Normalization technique	**Aggregate size of normalized series numbering**	**Average length of numbering**	**Increase over null routine (decrease)**
Null routine	4,881,946 characters	7.01 characters	—
System normalization	4,097,113 characters	5.88 characters	(16.12%)
1	9,147,581 characters	13.13 characters	87.30%
2	7,779,210 characters	11.17 characters	59.34%
3	8,130,532 characters	11.67 characters	66.53%
4	7,810,327 characters	11.21 characters	59.91%

System normalization reduces the size of numbering by removing punctuation—on average, about one character per instance. The four techniques described in this report increase the size of numbering by expanding numerals to a fixed length—adding, on average, about 7 characters.[134] Headings processed by techniques 2, 3 and 4 contain fewer characters than headings processed by technique 1 because these three techniques involve in some manner the removal of alphabetic characters from the numbering.

5 Conclusion

In the course of its investigations, the task group identified a range of approaches to the normalization of series numbering. These approaches vary in their sophistication, the correctness of the order of series headings they produce, the relative amount of time each requires to process a typical heading, and the amount of space occupied by the normalized series numbering. The use of any of these normalization techniques provides a better sort order for series headings than the order produced by no normalization at all, or by system normalization; the employment of even the simplest of these techniques carries with it a marked increase in the number of correctly sorted headings.

Approaches to the normalization of series numbering other than those described here might be devised. For example, there is a substantial gap between the complexity of the level 3 and level 4 techniques. It is likely that some parts of the level 4 technique could be grafted onto the level 3 technique, producing additional correctly sorted headings while also reducing the number of mislocations generated by the level 3 technique. Complex as it may be, even the most elaborate of the approaches to normalization is not necessarily complete. The testing of any sufficiently complex normalization scheme (such as the level 4 algorithm) against a different corpus of headings will almost certainly bring to light additional special cases for which allowance might be made.

[134] Or, to state this another way, by replacing one character with 8 characters, on average. When applying the instructions in Appendix C, all of the programs used for testing the normalization algorithms expanded numerals to a minimum length of 8 characters.

The task group makes no recommendation as to which solution might be the optimal one. Individual libraries and groups of libraries, working with the vendors of their automated library systems, will need to draw on the techniques described in this report, and similar techniques, to design a solution that best suits their needs and operating environment. The course to be settled on by library system vendors and their customers will depend on the perceived seriousness of the problem and the programming resources available. The task group encourages all parties concerned with the order of headings in library catalogs to consider the possibilities, and to implement some improved technique for normalizing series numbering. There is no doubt that library systems can in the future do a better job of arranging series headings than they have up to now, and that such an improvement need not come at high cost.

6 Compliance

The task group prepared a reference file from the corpus of series headings it used in its work. This file contains each series heading and its numbering in native form, together with the series numbering as normalized by each of the algorithms described in this report.[135]

The algorithms described in this report do not constitute a standard. Nonetheless, some value or merit may reside in the knowledge that a particular implementation of an algorithm for normalizing series numbering conforms to one or another of the algorithms. The task group approves the following standard for judging compliance with this report:

> A program that normalizes series numbering may be said to be in compliance with one of the algorithms described in this report if 100% of the normalized series numberings it produces match the corresponding data in the reference file.[136]

An implementation that passes this test may be described in promotional literature as conforming to level 1, 2, 3 or 4 of this report, as appropriate. Strategies for normalizing series numbering other than the four described in this report may be devised. Although such strategies may not be described as conforming to the letter of this report, if they result in a better arrangement of series headings they certainly conform to its spirit.

[135] Except as specified in this report, the data in the reference file are normalized according to the conventions of NACO normalization. For a period of at least 5 years from the date of issuance of this report, Northwestern University Library will undertake to make this reference file of series headings available to interested parties, for use in testing routines for the normalization of series numbering.

[136] Differences caused by the addition of words to one of the lists in Appendix D should not be counted as differences for the purpose of determining a program's compliance with one of the algorithms described in this report. Likewise, differences caused solely by the use of a normalization scheme for alphabetic characters other than NACO normalization should not be counted as differences for the purposes of measuring compliance.

Appendix A: Charge to the SCA Task Group on Series Numbering

Background:

In 1999, BIBCO (prompted by concern on the part of many PCC members about the display of series headings in library catalogs) appointed a Working Group on Series Numbering to consider changes to the presentation of series information in bibliographic and authority records and related matters. One of the problems that prompted the formation of this group was the inability of the current generation of library systems to arrange headings for series in order by the series numbering. (Systems arrange series subfield ǂv alphabetically, rather than numerically.) Among the group's recommendations are the following:

In June 2001, MARBI considered discussion paper 2001-06, which grew out of the report of the BIBCO Working Group. This discussion paper described the problem of the arrangement of series headings, and the current failure of library systems to solve this problem. The discussion paper presented a variety of solutions, most of which called for changes to MARC content designation and implied substantial retrospective conversion of existing machine-readable records.

In the discussion, there was general disfavor for any proposal that called for a change to MARC content designation. Instead, the opinion around the table was that the final solution proposed in the discussion paper—an algorithm for distinguishing captions from numbering that could be implemented by system vendors—should be explored before any change to MARC coding would be considered. It was generally recognized that such an algorithm would not and could not be perfect, able successfully to accommodate all cases; but that an algorithm that captured the majority of the common cases and resulted in no worse an arrangement for the remainder of headings than was already in effect would be good enough. MARBI returned the issue to PCC for the design of such an algorithm.

Charge:

The PCC Policy Committee charges the Standing Committee on Automation Task Group on Series Numbering to investigate ways in which local systems could provide improved displays of series headings, ignoring captions in subfield ǂv and arranging the numerical portions in numerical order. The outcome of the work of this group will be an algorithm for recognizing and ignoring captions in subfield ǂv that can be presented to library system vendors in a manner adaptable to each system. The algorithm will be as general as possible; its application to phrase searching, searches in which the numbering of a series is a secondary selection criterion, and aggregate displays of series headings, should all be considered. The report should describe in detail the categories of information found in subfield ǂv which are to be handled by the algorithm, and should provide a means for identifying those instances of subfield ǂv to which the algorithm should not be applied.

Time frame

A draft final report is due no later than the SCA meeting at ALA in Atlanta, with the final report to follow no later than the 2002 PCC Operations Committee meeting.

The final report should include the algorithm and supporting materials described above, and describe the work required on the part of library system vendors to implement it

Appendix B: Corpus of extracted headings

To obtain a set of access points for numbered series to be used in testing, a program passed through the Northwestern University Library database and found each bibliographic record that contained a series heading with subfield ǂv.[137] At the time of the program was run, Northwestern's database contained 3,178,122 bibliographic records; 672,389 of these records (21.16%) contained at least one numbered series heading. From these records 696,510 series access points were extracted. (The number of extracted access points is greater than the number of bibliographic records because some records contain more than one numbered series access point.) These extracted series fields form the corpus on which the task group's experiments were performed.

Tag	Number of occurrences	Cumulative percentage (n=696510)
400	456	0.07%
410	9427	1.35%
411	87	0.01%
440	324490	46.59%
800	6550	0.94%
810	84109	12.08%
811	438	0.06%
830	270953	38.90%

Not all of these series headings were well formed. For example, the subfield ǂv in a few headings contained only a full stop; a few headings contained subfield ǂv but no subfield ǂa. Headings in Northwestern's bibliographic records were corrected as these and other problems were discovered. Despite changes to the underlying bibliographic records, the corpus of series headings was not extracted repeatedly. The task group felt it would be better to work with a stable set of headings throughout the project than constantly to be shuttling updated files back and forth and re-performing tests. (The incorrect form of some headings in the corpus might also more fairly represent the state of the catalogs of individual libraries than would a perfectly tidy set of headings.)

The entries in the corpus represent 81,950 distinct series headings. The following table shows the number of occurrences for distinct headings that appear from 1 to 30 times in the corpus.

Number of times a heading occurs	Number of headings	Cumulative percentage (n=81950)
1	42573	51.95%

[137] Fields 400, 410, 411, 440, 800, 810, 811 and 830 were examined for subfield ǂv. The 490 field is only a *series statement*, not also a series *heading;* 490 fields were not considered. The Northwestern University Library database contains no instances of the obsolete 840 field.

2	12105	66.72%
3	6222	74.31%
4	3867	79.03%
5	2613	82.22%
6	1944	84.71%
7	1553	86.61%
8	1200	88.07%
9	969	89.26%
10	823	90.26%
11	682	91.09%
12	576	91.80%
13	491	92.40%
14	489	92.99%
15	366	93.44%
16	350	93.87%
17	280	94.21%
18	270	94.54%
19	242	94.83%
20	241	95.13%
21	206	95.38%
22	192	95.61%
23	169	95.82%
24	164	96.02%
25	160	96.21%
26	139	96.38%
27	128	96.54%
28	127	96.69%
29	97	96.81%
30	93	96.93%

At the other end of the scale, there are 35 headings represented by 1000 or more access points in the corpus. The most popular heading, *Early English books, 1641-1731* (the series heading for a microform set), is represented by 43,012 access points.

Appendix C: Handling of digits

Sequences of digits must be manipulated so that they may be sorted as values, not as characters. There are several ways to normalize digits to achieve this result. Perhaps the simplest of these involves the following elements:

- Left-justify with leading zeroes a contiguous sequence of digits within an area of some standard length. If a contiguous sequence of digits is longer than the standard length, remove any leading zeroes but otherwise leave the numerals alone.
- Precede the numeric sequence with a blank space (unless the numeric sequence is the first word in the series numbering)
- Follow the numeric sequence with a blank space (unless the numeric sequence is the final word in the series numbering)

The following examples (prepared according to the instructions in section 3.2) show the effect of this handling method on several instances of series numbering. For the purposes of illustration, a standard length of 8 characters is used here for each contiguous sequence of normalized digits.

Series numbering	**Normalized form**
`v. 15`	`V 00000015`
`1982, no. 16`	`00001982 NO 00000016`
`EPA-R5-73-017`	`EPA R 00000005 00000073 00000017`

The interesting question is: How many characters should be allocated to each sequence of normalized digits? The series numbering subfields in the corpus of series headings contain 912,846 contiguous sequences of digits. From the distribution of the lengths of these sequences shown in the following table, it appears that setting the minimum length of a set of normalized digits to either 6 or 8 characters would allow all but a trivial number of headings to be sorted correctly. (The examples included in this report all show a minimum length of 8 characters.)

Length of contiguous digits, minus any leading zeroes	**Number of segments of this length**	**Cumulative percentage (rounded; n=912,846)**
1	206,964	22.67%
2	333,360	59.19%
3	205,829	81.74%
4	101,612	92.87%
5	38,978	97.14%
6	26,058	100%
7	34	100%
8	8	100%
9	2	100%

10	1	100%

Adding zeroes at the left is not the only way to achieve the proper sorting of digits. For example, each sequence of numerals might be packed into a set of hexadecimal digits of some fixed length (so that the characters ‘286’ become 0x00000286); or converted into a hexadecimal numeral of some fixed length (so that the characters ‘286’ become 0x0000011E). Any representation of the value contained in a sequence of digits that allows it to be sorted by value should be acceptable; the choice of the specific technique to be used is a matter that may be settled by the system developer. In any case, the considerations mentioned above for determining the minimum length of the normalized sequence of digits remain in effect, *mutatis mutandis*. Leading and trailing spaces around the normalized digits must be supplied as described above.

The normalization techniques described in this report all assume that the machine collating sequence places digits before alphabetic characters. Library systems operating on machines whose collating sequence digits numerals *after* alphabetic characters—for example, the NOTIS system running on IBM mainframes, which use the EBCDIC collating sequence—may require the use of an additional adjustment to the numeric sequence, to force the digits to sort before other characters.

Appendix D: Lists of words used at various points

This appendix contains the lists of words referred to in the descriptions of level 3 and 4 normalization. It is likely that the application of one of these two normalization techniques to a set of series headings other than the corpus used by the task group for its testing will bring to light additional words of a similar nature that should be added to one or more of these lists.

D.1 Caption words

Single-character words: Single letters that may be used as abbreviations for caption words. The letter 'V' should not be included in this list.

D F H J N P R S T

Multi-character words: Words and their abbreviations that may be caption words. This list may include words with commonly-occurring typographical errors.

AARSSKRIFT AASTAK AASTAKAIK ABD ABDR ABDRUCK ABH ABHAND ABHANDLUNG ABHANDLUNGEN ABSCHN ABSCHNITT ABSTRACT ABT ABTEIL ABTEILUNG ABTH ABTHEIL ABTHEILUNG ADAD ADD ADDENDUM AFD AFDELING AFL AFLEVERING ALADAD ALBUM ALKITAB ALQISM ALSIFR AN AND ANEJO ANEXO ANN ANNEE ANNEX ANNO ANO ANY ARBEITSPAPIER ARG ARGANG ARGGRAFFIAD ARSSKRIFT ART ARTICLE AUSSTELLUNG AVD AVDELNING AVHANDLINGER AVIV

BAND BANDCHEN BANDE BD BDE BDCHEN BDCHN BEIHEFT BEIHEFTE BEILAG BEILAGEN BEITRAG BERICHT BHFT BIL BILANGAN BILDHEFT BIND BK BKS BLATT BLOCK BOOK BOOKLET BOOKS BROCHURE BUCH BUCHER BUL BULL BULLETIN

CAHIER CAHIERS CASSETTE CAT CATALOG CATALOGO CATALOGUE CATALOGUS CH CHAP CHAPBOOK CHAPITRE CHAPTER CHIH CIS CISLO COMUNICACAO CONFERENCE CONTRIBUTION CORPUS COURSE CUADERNO

DAI DEEL DEL DEELEN DELEN DIL DISC DOC DOCUMENT DOCUMENTO DOKUMENT DOSSIER DRUCK DZEL

EINZELAUSG ERGANZUNGSBAND ERGANZUNGSBD ERGANZUNGSHEFT ESTUDIO ESTUDIOS ETUDE ETUDES EXTRA

FASC FASCICOLO FASE FASCICLE FASCICULE FASZ FASZIKEL FG FICHE FICHES FILM FOL FOLJD FOLJDEN FOLGE FORUM FS

GRADE GRADES GRANTHANKAH GROUP GRUNNA GUIDE GUIDE-BOOK GUIDES GUIDEBOOK

HAFT HALB HALB-BD HALBBAND HALBBD HANDBOOK HANDLIST HAO HAUPTABTH HDBK HEFT HEFTE HELEK HF HFT HFTE HOV HOVERET

IMLEABHAR ISSUE ITEM ITEMS IWE

JAARG JAARGANG JAHR JAHRESREIHE JAHRG JAHRGAG JAHRGANG JG JHRG

KAN KAP KAPITEL KEREK KEREKH KITAB KN KNJ KNIGA KNIHA KOMPLEKT

LEAFLET LECTURE LEHRHEFT LEVEL LEVELS LEVNADSTECKNINGAR LFG LIBER LIBR LIBROS LIEFERUNG LIST LIVR LIVRAISON LIVRE

MANUAL MAP MAPPE MEMOIRE MEMORIA MICROCOPY MIS MODULE MONOGRAFIA MONOGRAFIE MONOGRAPH MONOGRAPHS MONOGRAPHY MUJALLAD

NIDE NIDE-T NIDOS NIVEL NIZ NO NOMBOR NOMOR NOS NOTE NR NRS NUM NUMBER NUMERO NUMERUS NUMMER

OBRA OEUVRE OPUS OSA

PAGE PAGES PAM PAMPHLET PAPER PAPERS PARATEMA PARS PART PARTE PARTES PARTIE PARTIES PARTS PERIODICAL PHASE PIRSUM POS POSITION PRAPARANDENHEFT PROGRAM PT PTE PTIE PTS PTES PUB PUBBLICAZIONE PUBBLICAZIONI PUBLICACI PUBLICACIO PUBLICACION PUBLICATION PUBLICATIONS PT

QISM QUADERNI QUADERNO

RADA RAEKKE RAPPORT RECUEIL REDA REEK REEKS REEL REELS REG REIHE REIRE REKKE RELATORIO RELEASE REP REPORT REPT RESEARCH ROC ROCNIK ROCZ ROCZNIK ROK RPT

SA SAGGI SARJA SATSU SAYI SAYISI SB SBORNIK SCHRIFT SEC SECT SECTIO SECTION SECTIONS SECTS SEFER SEGMENT SELECTIONS SER SERIE SERIES SERIIA SERIJA SESS SESSION SET SEZ SEZIONE SHEET SIFR SKRIFTER SKUPINA SONDERBAND SONDERBD SPECIAL STAGE STUCK STUDIA STUDIE STUDIEN STUDIES STUDIESTUK STUDY STUK SV SVAZEK SVAZOK SZ

TAGUNGSBD TAPE TEIL TEILBAND TEILBD TEXT TESTI TEXTE TEXTBAND TEXTS TH THEIL THEME THROUGH TI TITLE TITRE TITULO TL TOM TOME TOMO TOMOS TOMUS TOPIC TRACT TSUKAN

UNIT UNITS

VED VEREINSJAHR VERHANDELING VEROFFENTLICHUNG VEROFFENTLICHUNGEN VERSLAG VIP VO VOL VOLS VOLUME VOLUMEN VOLUMES VOLYM VORTRAG VORTRAGE VYP

WHOLE WORK WORKING WORKSHOP

YEAR

ZAHL ZEHNT ZESZ ZESZYT ZV

D.2 ‘Series’ words

Words that mean ‘series’.

COLLANA

DIVISION DIZI DIZISI

F FG FOL FOLDJDEN FOLJD FOLJDEN FOLGE

KOLO

P PERIODE
RADA REDA REEKS REIHE REKKE
S SARJA SER SERI SERIA SERIE SERIES SERIIA SERIJA SKUPINA

D.3 'New series' words

Multi-character single-word abbreviations for 'new series' in various languages. Note that the abbreviation 'NF' is not included here.

NFBD NS

D.4 'New' words

These words are only subject to manipulation when the preceding or following word is in the list of 'series' words.

ALTER ALTERA
N NEW NEUE NUEVA NIEUWE NOU NOUV NOUVA NOUVELLE NOVA
NOVAIA NUOV NUOVA NUWE NY

D.5 'Hors' words

These words are only subject to manipulation when the following word is in the list of 'series' words.

H HORS

D.6 Standard replacements

Under certain conditions, a normalization routine replaces information in an instance of series numbering with one of the following standard texts. Note that the standard replacement does not depend on the language of the original series heading. Note also that the standard replacements will properly cause 'new series' to sort before numbered subseries such as 'series 3'.

For 'new series': NEW SER
For 'hors series': HORS SER
For 'series': SER

D.7 Short words to be omitted from series abbreviations

& AN AND AT DE E FOR IN OF THE TO

D.8 Articles to be omitted from series abbreviations

A AN THE

D.9 Caption words to be replaced by a standard form in certain cases

Caption and other words to be replaced by a standard form when found in series numbering. Versions of a term in various languages are reduced to the same form, as are singular and plural forms.

Replace these words	**With**
AASTAKAIK	AASTAK
ABDR ABDRUCK	ABD
ABHAND ABHANDLUNG ABHANDLUNGEN	ABH
ABTEIL ABTEILUNG ABTH ABTHEIL ABTHEILUNG	ABT
ADDENDUM	ADD
AFDELING	AFD
AFLEVERING	AFL
ANEJO ANN ANNEE ANNO ANO	AN
ANEXO	ANNEX
ARGANG ARGGRAFFIAD	ARG
ARTICLE	ART
BAND BANDCHEN BANDE BDE BDCHEN BDCHN	BD
BEIHEFT BEIHEFTE	BHFT
BEILAGEN	BEILAG
BILANGAN	BIL
BOOK BOOKLET BOOKS	BK
BULL BULLETIN	BUL
CATALOG CATALOGUE	CAT
CONGRESS	CONG
DEEL DEL	D
DOCUMENT DOCUMENTO DOKUMENT	DOC
ERGANZUNGSBAND	ERGANZUNGSBD
FASCICOLO FASCICLE FASCICULE FASZ FASZIKEL	FASC
FICHES	FICHE
FOLGE	F
HAFT HEFT	HFT
HALBBAND HALBBANDE HALBBDE	HALBBD
ITEMS	ITEM
JAARGANG JAHRESREIHE JAHRG JAHRGAG JAHRGANG JAHR	JHRG
KNJ KNIGA KNIHA	KN
LEVELS	LEVEL
LIEFERUNG	LFG
LIVRE LIVRAISON	LIVR

MAPPE MAPS	MAP
MONOGRAFIA MONOGRAFIE	MONOGRAPH
NOMBOR NOMOR NOS NR NRS NUM NUMBER NUMERO NUMMER	NO
PAGES	PAGE
PAPERS	PAPER
PARS PART PARTE PARTES PARTIE PARTIES PARTS PTS PTES	PT
POSITION	POS
PUBLICACI PUBLICACION PUBLICATION	PUB
QUADERNI	QUADERNO
RAEKKE REEKS	REEK
REELS	REEL
REIRE	REIHE
RAPPORT REPORT REP RPT	REPT
ROCNIK ROCZ ROCZNIK ROK	ROC
SEC SECTION SECTIONS SECTS	SECT
SERIE SERIES SERIIA	SER
SESSION ZESZ	SESS
SONDERBAND	SONDERBD
ERGANZUNGSHEFT ERGANZUNGSHEFTE ERGANZUNGSHFT SUPLEMENTO SUPP SUPPL SUPPLEMENT SUPPLEMENTA SUPPLEMENTBAND SUPPLEMENTBD SUPPLEMENTO SUPPLEMENTARY SUPPLEMENTS SUPPLEMENTUM	SUP
TEIL TEILBAND TEILBD THEIL TOM TOME TOMO TOMUS	T
TEXTBAND TEXTS	TEXT
VEROFFENGLICHUNGEN	VEROFFENGLICHUNG
VOLS VOLUME VOLUMEN VOLUMES VOL	V

D.10 Roman numeral letters

I V X L C D M

D.11 Chinese, Japanese and Korean words

Words that constitute the first part of ordinal number labels

DAI DI TI

Words that constitute the second part of ordinal number labels

BU BUNSATSU CHI CHUNG FEN GO HAO HEN KAN KOZA NEN PIEN SATSU SEIJI SETSU SHU SOSHO TSE ZHONG

D.12 Beginnings of words that mean 'whole'

GANZ WHOL

D.13 Abbreviations that mean 'chapter'

ch. Kap.

D.14 Words that mean 'chapter'

CHAP. CHAPITRE CHAPTER KAPITEL

D.15 Words associated with ordinal numbers

Everything in the list of 'series' words, plus 'AN', 'ANNEE', 'CONG', 'CONGRES', 'CONGRESS', 'SESS', 'SESSION'

Appendix E: Method for testing series numbering

The task group explored the possibility that the series numbering example in an authority record (642 field) could be used by a program to assess the correctness of the form of the information in subfield ǂv of a bibliographic series access point. Such a test could be incorporated into automated library systems: If the two pieces of information were compared and found not to coincide, the library system could notify the cataloger of the discrepancy and the cataloger could take the appropriate action. By helping enforce consistency in numbering practice, the library system would indirectly improve the reliability of the sorting of series headings. Data gathered during the task group's explorations of this matter suggest that great value can be drawn from the use of the series numbering example to evaluate series numbering practice in bibliographic records.

The task group prepared a test program to compare the series numbering found in the access points in the corpus to the series numbering examples in the corresponding authority records. The program does its work by converting the series numbering example and the bibliographic subfield ǂv each into patterns, which it then compares. If the patterns match, the numbering is declared acceptable; if they do not match, the numbering is declared not acceptable.

As is the case with the preparation of the normalized form of series numbering for use in arranging headings, the design of the algorithm that reduces series numbering information to a pattern is a delicate matter; and here again, the more care spent on the design of the algorithm, the better the outcome is likely to be. Converting *digits* into a pattern seems to be a straightforward matter: the position of digits within the numbering is important, but the particular digits used in an instance of subfield ǂv are not. The handling of *captions* and other text appearing with digits is not so clear: text must be accounted for, but in what manner? Must the captions match exactly in all particulars of spacing, capitalization and punctuation, or is some degree of variation to be permitted? Can a scheme be developed that will raise an error if 'v.' is used instead of 'Bd.' but not allow 'no. (PHS)' to be used instead of 'no. (OHDS)'? Should there be one scheme for copy cataloging and another scheme for original cataloging—one to allow a heading that is 'good enough' to pass without calling for undue time-consuming changes, the other to enforce the highest level of adherence to the putative standard? The task force's experiments do not provide clear answers to all of these questions, but may indicate directions for further exploration.

The program used by the task group for its experiments employed as its foundation the standard normalization scheme referred to elsewhere in this report: NACO normalization. This choice reflects a carefully considered compromise. The use of some kind of normalization scheme during the derivation of a series numbering pattern allows a program (and therefore the cataloger assumed to be on the receiving end of error reports) to ignore minor points of spacing and punctuation and to concentrate on more critical matters such as the use of the proper caption. However, the use of any normalization scheme means that differences in spacing, punctuation and capitalization between the series numbering example and the bibliographic series numbering will go undetected, and therefore uncorrected; but these are differences not likely to be of great moment in any library system.[138]

The task group initially designed its test program with a single test of series numbering information. Although the program was basically correct in its handling of series numbering, many of the errors reported by the simple pattern-producing algorithm should not have been considered errors at all. To reduce the number of false reports, a second level of testing, employing a different pattern algorithm, was inserted at the point the program found a discrepancy in numbering pattern using the first scheme. This reduced the number of false error reports, but did not eliminate them entirely. It is possible that additional expansion in this manner of the scheme described here would further reduce the number of unnecessary error reports. (It is also likely that, given the current structure of the 642 field, no scheme will be able to eliminate false reports without also missing some conditions that should be reported as errors.)

The following is the core part of the logic used by the test program, presented in the same format used elsewhere in this report for condensed versions of program code. Note that this algorithm is couched not in the limited terms of the series headings included in the corpus (which by definition all contain subfield ǂv), but in terms of any series access point found in a bibliographic record, whether subfield ǂv is present or not; this algorithm could be used for any series access point that is represented by an authority record.[139] This algorithm is designed to detect not simply problems with series numbering in bibliographic records, but also problems in authority record coding.[140]

```
If the series numbering code[141] is 'a' (series is numbered)
    If the authority record contains a 642 field
        Inspect numbering information as described below[142]
```

[138] Given the benefits of the use of some normalization scheme, it might seem to be even more helpful if the system were to use the normalization scheme used for series numbering (such as one of the four normalization schemes described in this report) when it creates patterns from the 642 field and bibliographic subfield ǂv. The harmony between the two operations would prevent a given system from reporting discrepancies that don't actually make a difference in the context of that local system. This might well be the case, but this would also mean that records contributed to a shared database by users of various systems would reflect differences in practice caused by varying choices for series numbering normalizations by system vendors.

[139] The test program considered as 'series' authority records only those with code 'a', 'b' or 'z' in the type of series code (008 field, byte 12).

[140] At least some of the problems in authority record coding trapped by this algorithm could also be identified by a validation routine that compared values in one part of an authority record to values in other parts of the record.

[141] Byte 13 of the 008 field.

[142] Note the lack here of an explicit test for the presence of subfield ǂv in the series access point. If the series access point is not numbered, it will eventually be reported as being in error, because its numbering pattern, being empty, will not match any numbering pattern derived from the 642 field. Note also that the handling at this point of series access points found in *serial* records presents something of a dilemma. If the issues of a serial are separately numbered within the series, the access point will not contain subfield ǂv even if the series is a 'numbered' series; if the issues of a serial all bear the same number in the series, the access point will contain subfield ǂv. This means that it's just about impossible for a program to look at a series access point from a serial record and determine whether the absence of subfield ǂv is acceptable. Given the way in which the test program is written, series access points on serial records that quite properly

```
    Else
        A numbering problem exists (authority record indicates series is numbered, but no 642 field; possible error in authority record coding)
Else if the series numbering code is ‘c’ (series is sometimes numbered, sometimes not numbered)
    If the bibliographic series field contains subfield ǂv[143]
        If the authority record contains a 642 field
            Inspect numbering information as described below
        Else
            A numbering problem exists (authority record indicates series may be numbered, but no 642 field; possible error in authority record coding)
Else (authority record indicates series is not numbered)
    If the authority record contains a 642 field
        A numbering problem exists (authority record indicates series is not numbered, but contains series numbering example; possible error in authority record coding)
    Else if the bibliographic series field contains subfield ǂv
        A numbering problem exists (authority record indicates series should not be numbered, yet bibliographic record contains subfield ǂv)
```

Inspection of series numbering:

```
Convert both subfield ǂa of the 642 field and the series numbering from bibliographic subfield ǂv into a pattern in the following manner (here called ‘scheme A’). This pattern uses a zero to mark the location of digits.[144]
    Apply NACO normalization
    Replace each digit with the character ‘0’ (zero)
    Replace occurrences of ‘0 space 0’ with a single zero[145]
    Replace consecutive occurrences of the character ‘0’ with a single zero
    If the last 4 characters of the numbering are ‘space-ETC’
        Remove the last 4 characters
```

do not contain subfield ǂv would be reported as errors. Since the corpus of test headings was built from series access points that contain subfield ǂv, it contains no examples of unnumbered series access points, from records for either serials or monographs. The inspection of series numbering in serials is made even more complicated by the practice (illustrated in the *CONSER cataloging manual*) of including *partial* series numbering in some cases; for example, *DHHS publication ; ǂv no. (SSA)*.

[143] If the series numbering code is ‘c’ and the series access point does not contain subfield ǂv there is no error, as the lack of numbering in this case is acceptable, according to the authority record. There is no way for the program to ‘know’ that the cataloger forgot to include subfield ǂv.

[144] Examples of series numbering information (642 subfield ǂa and bibliographic subfield ǂv) reduced to pattern according to scheme A:

Original numbering information	**Corresponding scheme A pattern**
`20`	`0`
`no. 5`	`NO 0`
`8th, 1975`	`0TH 0`
`13-14`	`0`
`1, sup. 1`	`0 SUP 0`
`6. Bd., Nr. 5`	`0 BD NR 0`
`Bd. VIII`	`BD VIII`
`No. SSA-IM-85-22`	`NO SSA IM 0`
`971-972, etc.`	`0`

[145] Punctuation having been removed during NACO normalization, there is no need here for the more elaborate substitution tests performed by scheme B.

```
If the 642 subfield ǂa and bibliographic subfield ǂv as converted into patterns by scheme A do not match
    Convert both subfield ǂa of the 642 field and the series numbering from bibliographic subfield ǂv into a pattern in the following manner (here called ‘scheme B’). This scheme uses a zero to mark the location of digits and an ‘A’ to mark the location of those uppercase alphabetic characters that are not part of a word that contains lowercase characters.[146]
        If the last 5 characters of the numbering are ‘space ETC.’ (in any mixture of uppercase and lowercase characters)
            Remove the ‘ETC.’, the preceding space, and any comma that precedes the space
        Replace each uppercase alphabetic character that is not adjacent to a lowercase character with ‘A’
        Replace each digit with ‘0’
        Replace occurrences of ‘0 space 0’, ‘0 hyphen 0’, ‘0 comma space 0’, ‘0 comma 0’ and ‘0 full stop 0’ with a single zero[147]
        Replace consecutive occurrences of the character ‘A’ with a single ‘A’
        Replace consecutive occurrences of the digit ‘0’ with a single zero
        If the last 2 characters of the numbering are ‘0 full stop’ or ‘A full stop’
            Remove the trailing full stop[148]
    If the 642 subfield ǂa and bibliographic subfield ǂv as converted into patterns by scheme B do not match[149]
        A numbering problem exists (bibliographic pattern does not match authority pattern)
```

The test program was slightly more elaborate than this description might suggest, because an authority record may contain more than one 642 field. For example, it may contain multiple 642 fields if the pattern of numbering for the series has changed, or if different institutions follow different numbering practices. An institution may choose to mark the

[146] Note that this scheme does not incorporate many common aspects of normalization, such as the conversion of lowercase characters into their uppercase equivalents and the removal of punctuation. Examples of series numbering information (642 subfield ǂa and bibliographic subfield ǂv) reduced to pattern according to scheme B:

Original numbering information	Corresponding scheme B pattern
D-15	A-0
81C	0A
Bd. 3, Kapitel R.	Bd. 0, Kapitel A
12. Bd., 2. Abt.	0. Bd., 0. Abt.
vyp. D4-7	vyp. A0
80—B-5	0-A-0
CMS/CPE/582/83	A/A/0/0
Bd. VIII, etc.	Bd. A
no. SSA-IM-85-22	no. A-A-0

[147] Note the intentional lack of parallel manipulations for similar constructions containing ‘A’.

[148] Retain other terminal full stops, which are likely to be full stops associated with captions.

[149] Examples of series numbering declared acceptable as a result of the scheme B comparison:

Series numbering example (642 field)	Series numbering in subfield ǂv	Scheme B pattern
XXV, 4	XXIII, 3	A, 0
no. (OHDS) 84-30193	no. (SSA) 05-10375	no. (A) 0
ch. A	ch. J	ch. A
no. 09-MA-21	no. 23-ST-10	no. 0-A-0
A	K	A
4B	2F	0A

642 fields in its local copy of an authority record with its own code in subfield ǂ5. For example, catalogers at Northwestern University Library are instructed to add NUL's code to subfield ǂ5 of the 642 field that matches the numbering practice in an item being cataloged, and to add a new 642 field (with NUL's code in subfield ǂ5) if the item being cataloged reflects a numbering pattern not otherwise represented in a 642 field. Because the test program was designed to discover whether the 642 field could be used at all to test series numbering, it performed its comparisons with all of the 642 fields in each authority record until it either found a matching pattern or ran out of 642 fields. A real-world program, concerned with approving records being added to a particular institution's database, should consider only those 642 fields that are relevant in the local context.

The test program applied the algorithm described above to the 696,510 series access points in the corpus, which represent 81,950 distinct series headings.[150] The Northwestern University Library authority file[151] contains authority records for 55,988 of these distinct headings (68.32%), leaving 25,962 headings (31.68%) not represented by authority records. These authority records cover 605,542 of the access points in the corpus (86.94%), leaving 90,968 access points (13.06%) not covered by authority records.[152]

The program found 61,405 access points (10.14% of access points with authority records) that present a series numbering problem when compared according to the pattern produced by scheme A; of these, 49,584 (80.75% of error reports; 8.19% of all access points with authority records) still seemed to present a series numbering problem after comparison according to the pattern produced by scheme B. The program eventually declared 555,958 access points (91.81% of the access points covered by authority records) to be acceptable after the two levels of inspection.

A review of 1% of the access points declared acceptable by the program (5560 items)[153] revealed 1 access point (0.02%) to have been incorrectly reported as acceptable.[154] If this ratio of incorrectly-handled to correctly-handled headings holds true for the entire corpus,

[150] Because the test program searches a remote database to retrieve authority records, it is difficult to measure the time required to perform its work independent of factors such as latency and system response time.

[151] The NUL authority file includes a copy of the entire Library of Congress name authority file.

[152] The average series heading covered by an authority record is represented in Northwestern's database by 10.82 access points; the average series heading not covered by an authority record is represented by 3.50 access points. The difference is probably due to headings for a few series (such as *DHHS publication* and series found in microform analytics) with high frequency of occurrence in the former group, and many headings occurring only once (including many that no doubt represent errors) in the latter group. (A typical example of an erroneous heading is *DHHS publication ; no. ǂv (ADM) 81-825 (SP).* Note the misplacement of the subfield ǂv code, which prevents the heading from matching its authority record.)

[153] The standard used in this review to determine whether or not a report was correct may be expressed as follows: Is the authority record coded correctly, and, if so, does the information in the authority record correspond to the information in the bibliographic record? This is not the same as asking whether all information coincides perfectly.

[154] In a test of an earlier version of the comparison algorithm, 3 access points in a sample of similar size were found to have been incorrectly reported as acceptable. The number of these errors is clearly small, but just as clearly not zero.

the test program would improperly declare 100 access points (0.02% of headings in the corpus represented by authority records) to be correct.[155] These errors committed by the program are hidden errors, because they would never come to the attention of a cataloger.

A review of 10% of the access points still deemed unacceptable after review by scheme B (4959 items) found that 3815 (76.93%) represented true problems and 1144 (23.07%) were incorrectly reported as problems. (These 1144 represent overt errors—they would be reported to a cataloger, who would evaluate and then ignore them.) If this proportion holds true for all reports of errors, it means that the program erroneously reported 11,440 access points in the corpus as problems, and should instead have reported only about 38,144 errors.

Examples of series numbering correctly approved by the test program:

Series numbering example (642 field)	Series numbering in subfield ǂv
no. 160	no. 79
Vol. 1, nr. 1	vol. 1, nr. 6
5	72.
96th Congress, no. 22	97th Congress, no. 30
81C	79J
78-40	76-4.
AR/EUA/80-15	AR/EI/80-05.
1987	1982-1988.
84-A-1	88-C-005
F590	S1857
no. 3	no. 13, etc.
v. 156	[v. 187]
87-768 F	95-815 A
TMS-820-4	TMS-297
no. 27	no. 5/5-5/7
Nouv. sér., t. 10, fasc. 1	Nouv. sér., t. 22, fasc. 2.
47, no. 2-3	49, no. 5
no. (OHDS) 84-30193	no. (HRA) 82-131
1929.IX.8	1931.IX.22

Example of series numbering incorrectly approved by the test program:[156]

Series numbering example (642 field)	Series numbering in subfield ǂv
v. 1, no. 1, 1969-70	v. 4, no. 1, 1972/1973

[155] This and the remaining figures in this appendix that are derived by extrapolation should be understood to be surrounded by a certain degree of statistical fuzziness. Unfortunately, no one in the work group is qualified to characterize this fuzziness properly. Nonetheless, the general tendency should be clear enough.

[156] The numbering example and subfield ǂv both reduce to the same scheme B pattern: ‘v. 0, no. 0, 0’.

Examples of series numbering correctly reported as a problem by the test program:

Series numbering example (642 field)	Series numbering in subfield ǂv
no. 5	v. 10
module 5	model 15
Bd. 19	1
nouv. sér., no. 4	new ser., no. 4
3rd ser., v. 1	ser. 3, v. 3
v. 31, article 2	vol. 16[157]
ser. A, no. 1	sez. A., no. 1
no 642 field	4
#2	1
9	20th
n.s., no. 154	new ser., no. 909
45. Bd.	Bd. 64
Teil 1, Abt. 6	T. II, Abt. II, 1
DLC[158]	v. 2
v. 12	v. l2[159]
pt. 1	pts. 7-8
no. 8	N° 35

Examples of series numbering incorrectly reported as a problem by the test program:

Series numbering example (642 field)	Series numbering in subfield ǂv
9A	6
1st report	2nd report
H2/79/1	A4/77
Bd. 7	Bd. 9, T. 1
t. 3	t. 11, pars. 3
no. 11, rev. 1	no. 1
spring 1995	winter 1996
nouv. sér., 10	6
ch. B	ch. F-G
4th ser., 1	3rd ser., 4
5	96, section 202.23
map C-97-A	map C-142
BS8	AS241Y
8a	15b
1198 c/a	1555 a/a
v. 13, no. 2	v. 31, no. 2, suppl.
Bd. I/1	Bd. XVI/1b[160]

[157] This is a correctly reported problem because of the discrepancy between ‘v.’ and ‘vol.’

[158] The authority record is coded improperly.

[159] This is a problem because the fourth character in subfield ǂv is the letter ‘el’, not the numeral one.

Adding the 38,144 errors correctly detected by the program to the 100 errors missed by the program means that the program should have reported about 38,244 errors. This is 6.32% of the access points in the corpus represented by authority records. This means that about 93.68% of the access points in the corpus are numbered correctly.

Adding the 11,440 errors that should not have been reported after the scheme B comparison to the 100 headings that were incorrectly approved during the scheme B comparison gives a total of about 11,540 errors committed by the program, which is 1.91% of the access points covered by authority records. 99.13% of the errors committed by the test program were errors on the side of caution—they are reports of conditions that are not in fact problems. The test program handled 98.09% of the access points correctly.

The algorithm for testing series numbering described in this appendix could probably be made a trifle more elaborate with at least some small profit. It could for example be designed to deal with series numbers that include 'optional' designations such as 'rev.', 'suppl.', 'bis' and 'appendix'. However, given the present form of the 642 field, it is not clear that the effort required to develop a routine vastly more elaborate than that illustrated here would be repaid.[161] What is clear is that the implementation of this algorithm, or of some other algorithm designed to perform a similar comparison, could help improve the quality of series headings in bibliographic records.

The letter of transmittal for this report includes the suggestion that library systems should use the authority 642 field to enforce better consistency in series numbering practice, through an algorithm such as the one described in this appendix. The task force would also like to propose the following changes to practice (or affirmations of current practice not always strictly adhered to) regarding the use of the 642 field in authority records.

- There should a written statement describing what constitutes a distinct pattern of numbering,[162] and there should then be a 642 field for each distinct numbering pattern used with a series heading.

[160] The incorrect report here is caused by the presence of the lowercase 'b', not the roman numeral. The 642 field reduces (scheme B) to the pattern 'Bd. A/0', the subfield ǂv numbering to the pattern 'Bd. A/0b'.

[161] Were the 642 field to be redesigned better to accommodate the needs of automated verification of series numbering in bibliographic access points, then development of a more elaborate algorithm would be appropriate.

[162] NACO participants already apply NACO normalization when deciding whether or not a reference tracing should be added to an authority record in the shared authority file, without taking into account to the normalization scheme used in the local library system. Agreement on a parallel standard for the series numbering example would lead to greater uniformity of practice in shared bibliographic records.

```
130 ǂa Cambridge studies in medieval life and thought[163]
642 ǂa 4th ser., 1 ǂ5 ...
642 ǂa 3rd ser., v. 5 ǂ5 ...
642 ǂa new ser., v. 4 20 ǂ5 ...

130 ǂa Letteratura italiana. ǂp Storia e testi
642 ǂa v. 35 ǂ5 ...
642 ǂa v. 44, t. 5 ǂ5 ...

130 ǂa Early English books, 1641-1700
642 ǂa 755:13 ǂ5 ...
642 ǂa 228:E.2, no. 8 ǂ5 ...
```

- Practice in local institutions should be either to mark all relevant 642 fields in the local copy of authority records with their own code in subfield ǂ5, or show acceptance of all 642 fields in the authority record by not marking any of them with the local code.

- If any 642 fields in the authority record are marked with a code for the local institution, the local library system should in the process of testing bibliographic series numbering consider only those 642 fields; if none of the 642 fields in the authority record caries the local institution's code, the system should consider all of the 642 fields in its test. The system should declare bibliographic subfield ǂv to be acceptable if its numbering pattern corresponds to the pattern found in *any* of the relevant 642 fields. Because it is not possible for a program to use the 642 field as presently constructed to evaluate the contents of bibliographic subfield ǂv with complete reliability, the library system should never refuse a bibliographic record simply because a test of the contents of subfield ǂv fails.

[163] The numbered/unnumbered series code in this authority record is 'c' (sometimes numbered, sometimes not). The original group of items (before publication of the 'new series') were not numbered. Because of differences in lowercase characters, 'new ser.', '3rd ser.' and '4th ser.' all represent distinct numbering patterns. There would be no need for a 642 field describing the numbering for '5th ser.' (if it exists) if it follows the same pattern (as determined by application of the series of tests described in this appendix) as does '4th ser', namely '0th ser., v. 0'.