U.S. GOVERNMENT PRINTING OFFICE
KEEPING AMERICA INFORMED

Versioning: The Future of Managing Changes in Digital Content at the U.S. Government Printing Office

Contents:

I. BRIEFING TOPIC: Versioning: the Future of Managing Changes in Digital Content at the U.S. Government Printing Office

I.1. SETTING THE STAGE

Digital content changes on a very frequent basis, and it is easy to transform content without identification or notification to customers. Customers want to be certain that they are using the most up-to-date version of information, and also want to be able to track the history and/or any changes that have occurred in information that is available to them. In the case of Federal information, multiple versions of government publications are often publicly available on multiple Web sites, which can be confusing and sometimes damaging to customers who are not aware of the version status of the material. Version tracking is a necessary operation in Government information management businesses to ensure citizens are accessing the proper documents. The following facts are presented to assist in the scheduled discussion:

Broadly defined, a "version" is a particular representation or "concrete instance" of a publication, which may also represent a particular form or variation of an earlier or "original" publication. We define "versioning" as the managing (storage and retrieval) of versions of publications. In terms of a simple Open Archival Information System (OAIS) reference model, versioning of information falls under the larger umbrella of content management in its current form at GPO, along with the elements of authentication, access, and preservation. Our expectation is that this model will also be used in the development of a future system.

Recognizing the need for an enterprise wide solution, GPO is investigating the implementation of a Content Management System (CMS), but currently has multiple individual systems that do not integrate. Without a CMS in place, GPO has relied upon descriptive metadata files integrated into search and reference tools to control versions of information.

The proposed future system envisions a scenario under which version certification will be required from the content originator for publications made available to GPO whenever possible.
GPO will be responsible for managing and preserving all versions of information and establishing authenticity throughout the lifecycle of the publication. In conjunction with this work, it is recommended that GPO and other agencies take proactive steps to make customers better aware of the versions of information they are using, through the use of both technology and the implementation of guidelines or standards for identifying versions of government information resources. The standards setting process should involve the wide array of stakeholders who use government information.

I.2 MICRO RECAP

The focus of the topic of "versioning" involves tracking and controlling information content. We are currently all working under a deluge of information, particularly in the digital world, that changes very quickly. There is a need to manage all of these different versions of information, while ensuring customers that they are getting official content and the most current versions.
Likewise, there is a need to provide researchers and historians with the ability to track different versions through audit trails back in time to ensure that all documents in a workflow model are authentic and can be appropriately "versioned".

In the print environment, only the issuing agency absolutely can decide if a publication is a new version when they send it to GPO be re-printed. In the digital environment, GPO has the responsibility to preserve information for permanent public access but agencies are changing it constantly without officially issuing

new versions. In many cases we see websites where documents are changed, reposted with the same file name, and customers don't know they are looking at a different version. In this case, versioning integrates with GPO's respective authentication and preservation processes.

II. REVISED ASSUMPTIONS

II.1. GPO should acquire, authenticate, catalog, preserve, and provide access to all published versions of official and authentic electronic Government publications that have been approved for dissemination to the public and are within scope of the FDLP.

II.2. GPO needs to address the issue of version control in conjunction with authentication, harvesting digital content, and permanent public access.

II.3. GPO should work in conjunction with agency publishers on the issue of version control. For electronic publications that are marked "Draft" or that are marked "For Official Use Only" or "For Internal Use Only," GPO will coordinate with agency publishers on the issues of cataloging and access.

II. 4. Policies should be developed that define what constitutes a new version of an electronic publication. The policies should provide information on when a digital version becomes a distinctly different entity.

II.5. GPO will continue to follow national bibliographic standards but policies must be developed to address version control for different types of electronic resources (e.g. monographs, serials, series, databases, Web sites, video, audio, data streams, integrating resources, multimedia).

II.6. In an automated version control system, subjective evaluation and/or interpretation will be necessary to develop and monitor version triggers.

II.7. Version information will be stored in the metadata associated with a given publication. The metadata will include information about the version of the publication being described and alternate versions of the publication.

II.8. End-users want to be able to track the history and/or changes that have occurred in Government publications. This information will be available in the digital signature or other future means of authentication for a publication.

II.9. GPO may explore value added products that augment the attributes of a publication (e.g. structure, content, format) for the purpose of distinguishing one version from another.

III. QUESTIONS TO COUNCIL, WITH COUNCIL DISCUSSION

III.1 QUESTION: Do you agree addressing the versioning issue is something GPO clearly needs to do, in conjunction with authentication and permanent public access? Is this more of an issue for GPO, even though we are not publishers of the content, in the electronic age as opposed to the print age?

DISCUSSION BY COUNCIL

Yes, it is GPO's responsibility, particularly because of the management issues involved in the electronic collection. However, the publishing agency does not relinquish control over the material, and GPO needs to work with the Federal publishing agencies on this.

The Interagency Committee on Government Information is also working on metadata standards with executive agencies and other groups.

III.2 QUESTION: How do you define what constitutes a new version in the digital realm? In other words, when does the digital version become a distinctly different entity?

DISCUSSION BY COUNCIL

In spatial data mapping information, for example, the version of Florida in May is not the version of Florida that we have now, because beaches have eroded. How frequently do you update that kind of information? Is it once a day? Every 30 seconds? Is it once a week? We should not look at version control simply in terms of text files and their supplements, but also version control of things that are going to change without human interference. How are we going to keep control of the national map of the United States, for instance? We need to make sure that those types of data are on the table.

III.3 QUESTION: When Federal agencies are constantly changing information on web sites on a monthly, daily, or hourly basis, putting up new versions, and overwriting existing files, which ones should GPO download and save?

DISCUSSION BY COUNCIL

We do not have an answer. Maybe GPO needs to ask when do the differences become significant enough to call it another version.

In the past, we have had superseded items, and there were rules developed for keeping or discarding each version. There were titles that were superseded that some of us did not discard, because we have students who need the older information. It is a social issue that should be worked through with the library community, especially if you are going to supercede items.

In the best of all possible worlds, the master set that GPO is having published will have all of the different versions.

GPO should be very open and transparent about what decisions are made about the frequency in which databases are updated.

GPO should work with NARA in its efforts to capture official government information and preserve that content over time. GPO also needs to stay on top of the work being done by e-GOV Initiatives and by the National Digital Information Infrastructure and Preservation Program.

GPO should look at other initiatives and undertake some test projects to identify strategies for capturing different versions.

We need to set priorities and determine what is actually doable, in terms of saving documents monthly, daily or hourly. There are limits, and the rules and the policies we establish can help define them. The social problems and the technical problems are wired together. Technical problems seem easy to address in many ways, but the social or the policy problems, we cannot answer.

III.4 QUESTION: GPO envisions using a content management system with associated parameters identified for version triggers, and their activation will result in version detection. What are your thoughts on this? Do you think that subjective human elements, once identified as setting parameters, can result in a triggering and activation process that is entirely or almost entirely automated? How much automation is possible in this process?

DISCUSSION BY COUNCIL

In an automated system there is a complex of factors involved, based on rules that are initiated by analysis and by thinking through what constitutes new versions. Once the process is automated, the human subjective intervention takes place, detecting new issues for versions and implementing new rules to reflect those issues.

.

You have to have rules and procedures that are performed by whatever artificial intelligence the system can deliver, but on top of that you need quality control that can tell you how trustworthy your original perceptions were. Artificial intelligence will get better and more complex, and will be able to make more complex decisions. But you need to know if the system can be trusted to do what you want, and you need to have a way to check that.

The metadata can help a lot with the automated triggering of file changes. For example, the metadata standard Dublin Core has a review date that could be included on an agency website, to be used as a trigger. If your metadata is richer, if you have more than one substantive element that you can compare to another substantive element of metadata, then you might have something that is a bit more than merely a trigger, and that would activate human intervention.

It is dangerous to develop a system that doesn't include the subjective human element.

III.5 QUESTION: What role do you see metadata playing in terms of reflecting versions and the relationship between e-publications and other publications?

DISCUSSION BY COUNCIL
A subjective human role; there's going to be a lot of interplay.

The metadata is going to be very useful in identifying different versions. If the metadata does not show the date of creation, and possibly show what the change from the previous version was, a person would have to compare each one word by word to find out what the difference was. For example, in a simple policy report, when you know that there is some sort of change, the metadata should indicate the substance of the change.

III.6 QUESTION: What role do you see for the use of authentication processes such as digital signatures in helping to maintain version control?

DISCUSSION BY COUNCIL
Authentication through digital signatures will become harder as different versions are created. It is easier to do for the first version.

Date stamping along with the authentication of the signature, and marking the most recent version with some unique identifier, would let people know if they were looking at the most current version or an older one.

One thing to consider is the technology changing over time. There have been multiple kinds of digital signatures; they change, and sometimes people crack the code.

III.7 QUESTION: What other issues have we missed that we should be thinking and talking about?

DISCUSSION BY COUNCIL
We need to consider how to handle the so-called loose-leaf services that issue intermittent replacement pages.

The agency websites themselves are an object of interest. We tend to focus on the individual documents, but the website itself as a construct could also be captured at specified intervals.

We do not want to lose sight of the fugitive documents issue. A current version of a fugitive document would be better than not having a version at all.

The legal issue will be an important one, when the legal community needs to recreate an authentic version of a document as it existed at a particular point in time.

IV. AUDIENCE QUESTIONS ADDRESSED DURING THE MEETING

The facilitator of the Council sessions accepted questions from the audience written on GPO-supplied cards. Eleven of thirteen questions were answered during the Council session. Those questions and their answers are summarized below. Two questions held to answer at a later date, either because of time constraints or the need for a subject matter specialist to provide a more detailed answer, follows the questions answered during the session.

IV.1 **QUESTION:** So far all of the responsibility is "GPO should". What is the depository library community's role in version control? Active? Passive?

RESPONSE: The role of the depository community is absolutely vital to this whole process. We have to have extensive feedback from the users, both the agency users and the end users of the content. They will help with identifying what is working or not working. There is an enormous value to having all of those eyes and ears out there looking for things that we might have missed, or identifying places where the rules may not be functioning and need to be tweaked. The relationship continues to be an extraordinarily important partnership in terms of making the system work.

IV.2 **QUESTION:** Are the needs of the originating agency the sole determining factor? Are there legitimate uses/needs by citizens and scholars for earlier versions even in those circumstances when the originating agency has no perception of value in earlier versions because the internal purpose for that document focuses only on the latest versions?

RESPONSE: No, the agency's needs are not the only need. And yes, meeting the users' needs is one of the reasons why the depository program has been so effective, and we will continue to take the public needs into consideration.

IV.3 **QUESTION:** When might it be necessary to retain uncorrected versions after correction is issued? Typographic corrections that do not change the meaning of a sentence may or may not necessitate retention of an earlier version, but typographic corrections to data in a table probably require retained access to incorrect data.

RESPONSE: That is where those subjective human elements come into play.

IV.4 **QUESTION:** In the various versions do they contain structures edited into different colors so people can easily identify the edited portion of the document?

RESPONSE: Not unless an agency is generating it that way. In Congressional bills you will often see strike through and inserted text where you can see what the editorial changes are, because the bill is a living evolving document and that is partly how that is communicated. It might be prohibitively expensive to do now, but would be something to keep in mind for the future.

IV.5 **QUESTION:** We talk about metadata for electronic versions and marked tangible versions in the ILS. Will there be a system in place to eliminate duplicate work in creating metadata and marked records?

RESPONSE: That is where XML is tremendously helpful, because it is becoming more of a universal standard in data interchange. The issue is that sometimes the information we receive is incomplete, causing a reliability issue in the system. It would be desirable to define guidelines or standards for what metadata elements are necessary to capture.

The tools need to be there for the next generation system design. The publishing agency can contribute or deposit the information with the appropriate metadata. We need to centralize and reduce redundancy in creating records for different programs. We need to be sure that we are capturing the information once and using it for multiple purposes, so that we have records that can flow into other metadata.

IV.6 **QUESTION:** Which government information might require very frequent data change, like stock market prices? For most users, the opening, closing, high and low figures for the day are sufficient, but a researcher tracking impacts may need frequency of at least a quarter of an hour. Are there specific parallels in Government information that we can identify? Stream flow data is already collected and retained as real time data. Can the agencies assist in determining other data that may fall into this category?

RESPONSE: Whether the agency keeps the data minute by minute would be determined by whether the agency has a purpose for getting the information that frequently.

It is an agency decision as to what they keep, and it would be unrealistic to think that GPO is going to ever have that kind of storage space. But there are beginning to be some models for doing that kind of capture on a consistent basis.

The publishing agency, GPO and NARA have to be working very closely together to determine what the agency is keeping for its own purposes, what it is required to keep under the Federal Records Act, and what might be of interest in terms of public use and educational value.

IV.7 **QUESTION**: (For Mr. Hemphill) Do you store anything other than monograph type documents, for example, serials or databases?

RESPONSE: Yes, we do. We go out and capture chains of information on the web, which is just one example. As far as multiple types of information, we capture black and white, color, two sided, landscape, portrait, all kinds of renditions of things.

IV.8 **QUESTION:** (For Mr. Hemphill) Do you keep the superseded documents? Some of these can have significant value to the historians.

RESPONSE: Yes, we do, where possible. We are one of the go-to places for historical information.

IV.9 **QUESTION:** An earlier question was misunderstood. It was asked if the collection will incorporate dynamic data files or was it also created on the fly. This is an issue of versioning. An example of this that might help is American Fact Finder.

RESPONSE: Our intention is to be as comprehensive as possible and to work with agencies on these kinds of things.

IV.10 **QUESTION:** Would it be possible for GPO to publish a concise abstract on the National Bibliography and Collection of Last Resort following a scientific method model list for bringing about the outcome? I realize this is a work in progress but something that is very concise needs to out there for all of us to understand.

RESPONSE: We are expecting to do substantial revisions of the various documents that we have been discussing and to begin consolidating them into a single plan instead of looking at this as individual facets, so we can certainly take that into account as we rewrite them. One of the whole values of this meeting was to take all these documents that have been developed at different point in time and begin to harmonize the documents and recognize that things are going to change.

IV.11 **QUESTION:** Will the different versions of the current GPO plans be archived and physically or publicly available? Will these Council briefs and the assumptions and the questions be posted so that others can see them?

RESPONSE: Yes. We have had a plan out for comment and then it has been revised. We have the more current version of all the reports on the main page with a link back to the older one so they can be compared. The assumptions came out of the older documents, and revised assumptions will be stated in the new documents.

V. AUDIENCE QUESTIONS ADDRESSED AFTER THE MEETING

V.1 **QUESTION:** Has GPO considered the use of a service mark for those online titles that are "Officially Official"? Such a mark could be used as a label on web pages and search results in engines such as Google--maybe a circled letter "O".

RESPONSE: In the near term, GPO's implementation of the Public Key Infrastructure digital signature solution will be the indicator of official status.

V.2 **QUESTION:** How do we determine the granularity of versioning for born-digital items, and how do we know what makes a new version?

RESPONSE: GPO is in the process of drafting a white paper outlining the policy issues surrounding version control, version triggers, and chain of responsibility for electronic resources that fall within the scope of the FDLP. GPO will solicit user feedback for best practices as we develop our definitional framework. This framework will build upon existing processes and standards currently employed by GPO and the library community.