

Organization and electronic management of parliamentary data and documents: intranet and databases

by Calogero Salamone

Summary

The Chamber of Deputies Library has been engaged for some time on the organization of the data on the internet and intranet websites of the Chamber of Deputies, and has carried out a significant amount of work relating to data acquisition and management.

Thanks to their professional qualifications and experience, librarians are ideally equipped to contribute to the planning, formation and management of data and data archives.

This paper makes a number of observations regarding new policies for the organization of parliamentary data on the website of the Chamber of Deputies, and examines several issues concerning the professional operators (librarians) and methodologies (standards and open formats, XML data markup) for the effective management of the data.

Organization of parliamentary data

For several months the Chamber of Deputies has been working on the creation of a new unified architecture that will arrange parliamentary documentation and data into an orderly system of information that is semantically structured, navigable, retrievable for consultation, easily accessible to different users and functional for diverse levels of research depth or specialization.

In keeping with the principle of open access, the new structure deploys open standards and introduces a new method of integration and coordination among the various departments of the Chamber, each of which can add their piece to the general “mosaic” of services. To this end, they may adopt shared models or design new guidelines and templates for research methods or else organize documents, materials and data into useful categories.

The building of a centralized framework and the adoption of a modular approach to the compilation of documentation will not only create a new and more effective tool for internal use, but will also provide external users with a comprehensive and transparent guide, enabling them to access the large mass of materials contained on the site. In this way, the processes by which important public policies are put into effect will be elucidated. Users can map out their own research paths by combining and recombining documents, data and information in accordance with their needs.

Greater concision and better organization, including in the layout of graphics, will facilitate the comprehension of complex themes and convey the essence of parliamentary activities, with the result that processes for the framing of policies and making of laws will become more transparent.

When the project to introduce a new methodology for the management of parliamentary data advances to its next stage, in which the experience and functional skills acquired by the Chamber of Deputies Library in the management of parliamentary databases can be leveraged, a new parliamentary information system involving both the Upper and Lower Houses can then be built. The possibility of this coming into being will be enhanced if the management of parliamentary data is unified from the moment data is extracted and databases are populated. Such a system would do away with all unnecessary duplication and inconsistencies of data management, and allow unified and integrated access to all information pertaining to parliamentary activities.

Data management and the role of librarians

By leveraging the resources of the Office of Parliamentary Databases and assigning all documents relating to policy-setting and oversight to appropriate categories using the EUROVOC thesaurus,

\the Library is helping to re-map the information paths. The work is now being done also by means of advanced techniques of semi-automatic classification, which use applications that analyze natural language and suggest possible classification descriptors that the operator may then choose to adopt.

The Library's normal sort of contribution, however, consists in the substantial task of mining data and entering information in the general database for documents relating to policy-setting and oversight, which encompasses all the documents of the Chamber of Deputies and Senate and all stages of their consideration by Parliament. The Library also manages a second large database that systematically records all the activities of MPs, both on the Floor of the House and in Committees.

The Database Office carries out hundreds of thousands of operations on parliamentary data every year. The operations include the registration, validation, quality control, and verification of published data. The work is taxing and often complicated by the possibility of errors in the data flow and the risk that bad data and omissions may become permanently nested within the layers of published information. Once errors become entrenched, it is very difficult to pinpoint them in the myriad of published data, which simply cannot be continually reviewed and re-checked.

As the amount of available texts continues to increase, work processes accelerate and electronic storage capacity expands, we are faced with an ever greater risk of dispersion, redundancy and fragmentation or, in a word, chaos.

Bringing order to this potential chaos requires the collaboration of many different professionals from several fields.

As regards librarians, they are now acknowledged, in Italy as in other countries, as having a well-defined role to play in projects for the digitization of historical publications, print catalogues and data archives.

Another separate role, that of the "data librarian", has now become recognised and well established.

The data librarian has been around for several years now, and the origins of the profession date back to the days of microforms and magnetic media such as reel recordings and cassette tapes. The job description has evolved along with the technology, and now refers to librarians who deal with data in the broad sense of the term, and whose tasks therefore include collecting, collating and managing electronic resources (including databases), dealing with 'suppliers' maintaining these resources in good working order, providing assistance for researchers and overseeing the digital preservation of information.

Several studies¹ have highlighted how libraries have the potential to promote the integration and reuse of data, and concluded that the essential task of the data librarian is to make the data fully and permanently accessible to users. Discharging this duty requires knowledge of several fields relating to formats, standards and metadata², all of which demand a high degree of collaboration and interaction with ICT departments.

Yet it remains difficult to pin down precisely what the tasks of data librarian are.

-
- 1 See, for example, Anna Gold, Massachusetts Institute of Technology, *Libraries and the data Challenge: Roles and Actions for Libraries*, in *Cyberinfrastructure, Data, and Libraries*, Part 2, D-Lib Magazine, September/October 2007, Volume 13, n. 9-10 (<http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>); and her paper *The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs*. Report to the JISC, July 2008, prepared by Alma Swan and Sheridan Brown, Key perspectives Ltd (<http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf>)
 - 2 On the role of the "data librarian", see Stuart Macdonald, Luis Martinez Uribe, *Data librarianship, a gap in the market*, CILIP, May 2008 (<http://www.cilip.org.uk/publications/updatemagazine/archive/archive2008/june/Interview+with+Macdonald+and+Martinez-Uribe.htm>)

In fact, when data is being created, the system design, the choice of format, the standards to be applied, the configuration of the metadata, the software used, the levels of accessibility and the search and retrieval functions are factors that will determine all the subsequent processes, including those assigned to the data librarian.³

Data librarians therefore have good reason to become involved as far upstream as possible in the process of creating and organizing data. Suitable intervention during the data modelling phase is therefore a particularly effective way of facilitating the librarian's later function of disseminating, using and preserving the data. Data that has been created using open and lasting formats and standards is far easier to preserve access and re-use.

Even so, as the literature for the sector often has occasion to note, data creation and management is a field that still lacks a well-defined and recognised professional qualifications based on specific training and instruction. What happens, therefore, is that persons with different roles and functions are taken on with reference to the experience and knowledge they have accumulated on the ground. Certain recent studies, such as that conducted by the Joint Information Systems Committee (JISC)⁴, have attempted to discriminate between the different functions and proposed that new professional roles be officially established.

On the other hand, some of the contributions that librarians can make have been quite clearly defined already⁵, and several studies have set out the scope and possible future developments of the librarian's functions⁶, arguing, for instance, that "by collaborating closely, and early, in the research process, librarians may become involved in creating data curation prototypes, or otherwise supporting the use of documentation, practices, or standards that will assure the longevity of the data downstream."⁷

The Chamber of Deputies, too, has yet to make a clear demarcation in the roles of those involved in the management of data. It is not entirely clear, for example, at what point the job of the librarian ends and the job of the IT professional begins, with the result that many tasks are assigned and done on the basis of personal experience and habitual practice.

The Database Office follows just such a hybrid approach in its activities, which encompass such diverse functions as librarianship – its starting point -, data management, the input of information into the parliamentary databases, the development of IT programs and even the modelling and application of standards. The Office has always operated in contact with the IT Department, with which it has very close ties of cooperation and exchange, also with different standpoints. In our

3 Margaret Henty, *Dreaming of data: the library's role in supporting e-research and data management*, Australian Library and Information Association Biennial Conference, 2-5 September 2008, p. 2 (http://www.apsr.edu.au/presentations/henty_alia_08.pdf)

4 *The skills, role and career ...*, 2008

5 *The skills, role and career ...*, 2008, p. 24: "The library and information science community should have an important role to play in the data science arena, particularly in delivering awareness and understanding of data issues and the importance of good data science and data curation. There are generic data handling and management skills that are native to librarians and can be taught as part of the basic research skills training in an institution. After all, the fundamentals of data science can be taught and subject expertise can be acquired over time."

6 Gold, *Libraries and data challenge ...*, 2007: "... librarians have a deep vested interest in systems of data publishing. Creating such systems will require new skills in including managing datasets as complex objects; it will also require creating standards related to data publication, description, citation, discovery, and reuse; and addressing policies for data disclosure.

At the most fundamental level, engaging the library profession in the problem of data management may lead to reframing the values and practices of the library profession. Where today library practices appear to be rooted in the management and delivery of objects (whether virtual or physical), from another point of view those practices are rooted rather in the management and "delivery" of relationships. And data is, after all, an encoding of relationships in the world ..."

7 *Ibidem*

case, the librarians' functions appear to have shifted significantly upstream towards the original data sources, even to the modelling of the systems to be used for the creation of the parliamentary data. This is what has been done in the case of, for example, the markup of parliamentary reports.

Unfortunately, the idea that librarians should be deeply involved in this area of activity is very often greeted with scepticism from two groups. The first group of sceptics are the IT people, who tend to believe that everything can be done (and done better) by automating all processes using the right application and by compiling an appropriate parser. The second are librarians themselves, some of whom subscribe to the idea that data management is the job of the IT department and has nothing to do with librarianship, which should concern itself with completely different matters.

Our experience enables us to us to say quite categorically that these two positions are both untenable. Above all, we have compelling evidence that adherence to either position runs the very real risk of creating serious gaps in the data, and, in particular, of compromising the standardization, accuracy and completeness of the data and thereby impairing the possibilities for its use and reuse. For this reason, librarians, who have a wealth of knowledge and experience and a keen awareness of the importance of and the need for standardization as well as a day-by-day experience of supporting research, must be involved in data management.

The original purpose of a library is to make its store of knowledge available and retrievable, to facilitate the use of the knowledge, to further its dissemination and to promote principles of sharing and collaboration.

Models and standards: Markup language

One of the problems that crops up in the management of the data of complex organizations, be it the Chamber of Deputies or a private corporation, is duplication. Processing the same data more than once increases costs, heightens the risk that the data itself will be lost and generates considerable difficulties when it comes to using the data, whether internally or externally for research and dissemination purposes.

Paradoxically, one of the biggest inconveniences of duplication is that as a result of its being processed several times, the data will often end being more difficult to find, because what is found may be incomplete, corrupt or out of date. Apart from cases where the search tools are not fit for purpose, the usual cause of data errors is the presence of anomalies in the processing at source. The information may not be recognised, or it may be corrupted or it may be incomplete.

When it comes to parliamentary data, there is an additional and compelling public interest in making sure the standards used are proper for the processing, sharing and reuse of the information, not only because parliamentary data is generated using public funds⁸, but also and especially because the data relates to essential knowledge that is public by definition. In this respect, ease of access bears an important relationship with the principle of democratic participation.

The fundamental question, therefore, is this: how can I make sure my data is always safe and findable?

The short answer is very simple. If we need to be able to find something again, we add a label, tag or mark of some sort to it. The marks will indicate that, say, this piece of text is a title, that other piece is a name, and so on.

We make markups every day when preparing texts. When we start a new paragraph, or put words in bold or italics or use inverted commas, we are assigning particular values and meanings to sections of text and marking out their different function with respect to other sections. The difficulty is that many of these marks and signs imply rather than state meaning: their significance is not immediately understood, nor can they always be defined in the same way. For instance, bold

8 V. M. Henty makes a similar point in *Dreaming of data ...* 2008

highlights a piece of text (but what textual function is it highlighting?) Italics may indicate a title, but may also indicate many other things besides.

The explicit markup of texts and, especially, the use of the XML standard (regulated by the W3C, the consortium that sets the rules and standards for the World Wide Web)⁹ makes it possible to standardize these practices.

For every type of document, then, a standard model must be selected and a set of rules defined and declared. The set of rules must then be scrupulously followed, and the documents validated to check for consistency with the rules. If the document fails to comply, the anomaly is noted and must then be corrected.

Markup language makes it possible to tag and name every data element from its first appearance, so that it can always be identified with certainty thereafter. In this way, we avoid duplication and can re-use and share data. Tagged data is easy to find, easy to systemize and easy to store and preserve over time.

Also, it costs less.

We must not miss the opportunity that this methodology offers. It is especially important to apply it at the early planning stages of a new information system, for it is far simpler to set these things up at the very beginning than to have to deal later on with incorrect or imperfect practices.

⁹ W3C Consortium, <http://www.w3.org/XML/>