

DIGITAL SENSE AND NONSENSE: DIGITAL DECISION MAKING IN SOUND AND AUDIOVISUAL COLLECTIONS

Ute Schwens, Deutsche National Bibliothek (German National Library)

Keynote speech at the 42nd IASA annual conference in Frankfurt, September 3rd to 8th, 2011

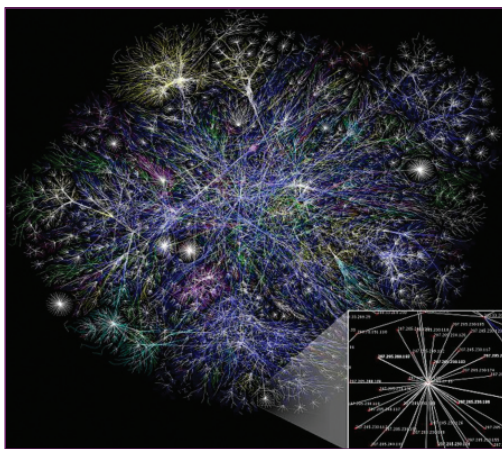


Figure 1: Visualization from the Opte Project of the various routes through a portion of the Internet

“The Internet is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless and optical networking technologies. The Internet can also be defined as a worldwide interconnection of computers and computer networks that facilitate the sharing or exchange of information among users. The Internet carries a vast range of information resources and services, such as the inter-linked hypertext documents of the World Wide Web (www) and the infrastructure to support electronic mail.

Most traditional communications media including telephone, music, film, and television are reshaped or redefined by the Internet, giving birth to new services such as Voice over Internet Protocol (VoIP) and Internet Protocol Television (IPTV).

Newspaper, book and other print publishing are adapting to Web site technology, or are reshaped into blogging and web feeds. The Internet has enabled or accelerated new forms of human interactions through instant messaging, Internet forums, and social networking.”¹

“The Internet can now be accessed almost anywhere by numerous means, especially through mobile Internet devices. ... Within the limitations imposed by small screens and other limited facilities of such pocket-sized devices, services of the Internet, ... may be available.

Service providers may restrict the services offered, and wireless data transmission charges may be significantly higher than other access methods.”²

1 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

2 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

“The low cost and nearly instantaneous sharing of ideas, knowledge, and skills has made collaborative work dramatically easier; with the help of collaborative software. Not only can a group cheaply communicate and share ideas, but the wide reach of the Internet allows such groups to easily form.”³

“World Wide Web browser software ... lets users navigate from one web page to another via hyperlinks embedded in the documents. These documents may also contain any combination of computer data, including graphics, sounds, text, video, multimedia and interactive content including games, office applications and scientific demonstrations. Through keyword-driven Internet research using search engines like Yahoo! and Google, users worldwide have easy, instant access to a vast and diverse amount of online information. Compared to printed encyclopedias and traditional libraries, the World Wide Web has enabled the decentralization of information.”⁴

Future visions? No, these statements are snippets from the most current definition of ‘Internet’ I found — at Wikipedia. Are we ready for using or serving these pictures?

I guess that most of us / most of our institutions have a big fund of data — metadata and even content — in our databases. Are these data open to everyone and open for being linked to other databases as it is requested in that definition of the internet? Do we use information from outside to enrich our own data? Do we foster our efforts to digitize objects that have been analog so far? Some of us do — and others are still very traditional, pointing out that e.g. there is not enough money to implement new processes and services. But moving institutions from the state-of-the-art to new thinking is not only a question of financial means. It is mainly a question of policy.

Policy in the context of digital media and the Internet includes many aspects of argumentation — among them the questions of

- Selection (Which types of objects should be digitized? For what reason? What is meant by ‘quality’ in this context?)
- Access (Open access vs. IPR? Material of poor quality? Material that can be found only once? Different, new access points?)
- Search and retrieve (Standards, networking, automatic indexing?)
- Re-use of data (Licensing? Commercial/non-commercial? Legal framework?)

I would like to give some examples for sensible activities in this context, bearing in mind that decisions often are dependent on formal conditions. Not always, however, not exclusively and not if professional answers have to be given.

Let me start with the selection of objects.

Do we need to digitize everything? Or do we have to keep our capability of valuing the meaning of objects for future generations? According to which criteria? I often discuss these questions with friends and colleagues — questions which are not new, but have been discussed continually over the last decades. When the German National Library started its collection of comics at the beginning of the 1950s, the whole library and scientific world was laughing or mocking about the ‘level of content’ in our magazines. And today? I no longer need to explain this.

Our collection at the German Exile Archive 1933 – 1945 contains documents, letters, photographs, manuscripts, etc. of Germans who had to emigrate during the Second World War. Imagine if we had kept only the personal papers of famous, well-known people! Or if we had thrown away all the material not directly related to the professional works of the persons!

3 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

4 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

And, especially when we are looking to our physical holdings, traditional media in our archives, libraries and museums have been collected over the years according to certain rules, collection or selection criteria, in order to fulfill the special tasks of the particular institution.

Digitizing these media means you do not have to do a second selection!

Therefore, looking to our traditional media, digitization is not a question of selection according to certain criteria, but according to financial resources, time line, priorities, quality and/or aims we want to reach. One priority for digitization could be the point-of-view of preserving the originals due to their poor condition. Another one could be the fact that there is only one existing specimen of a certain object. In both cases digitization is essential for usage: either because the original cannot be used any longer, or the original can be found only once — the digital version is accessible at any place and any time. I will come to this again later.

In this context the next decision to make is for the quality of digital objects. We should try to digitize as well as possible, but what does that mean? Should we produce a new product from the old material keeping only the content itself without the accompanying aspects, such as disturbing noises within sound files, little breakouts in films, dark paper in text documents, etc.? It is much more authentic and original to keep all these things together with the digitized object, in order to give users an idea of the original media. Alternatively you decide to create a new product, using the possibilities of modern digitization technology.

My own opinion on this is:

You should digitize as originally as possible, in order to preserve authenticity! Creating new products is on the producers.

I now come to a second aspect of digitization: the aspect of access.

As I mentioned already, through digitization activities, objects often are accessible again that could not be used before because of their poor physical, fragile conditions. The digitization itself of this material is a means both of preserving the originals (I mentioned this already as an aspect of selection) and offering the possibility to work with the content again.



Figure 4 and 5: Examples for the digitization of old piano cylinders

Perhaps you have already heard of our project at the Deutsche National Bibliothek (German National Library) to digitize the old piano cylinders from the end of the 19th/beginning of the 20th century. Some of them were made of wax and could not be given to users at all. A similar situation can be seen with piano rolls — made of paper and being used only with very special pianos, they cannot be given to users either. Three years ago the Music Archive of the German National Library digitized some copyright-free pieces of music that could be found on piano rolls as well as on shellac disc. Now access to this material is possible again for users.



Figure 6: CD, published by the German National Library, German Music Archive, in 2009

At this point I would like to emphasize that I am aware of the fact that access to these objects primarily is dependent upon copyright regulations. But supposing we could deal with and digitize copyright-free material, what is the best decision concerning access to this material?

1. Free access to everybody interested in the objects, or
2. Access restricted to defined user groups e.g. by financial, professional or even local criteria?

Even if digital objects are copyright-protected and therefore access via the Internet normally not allowed, we can offer a lot of information on digital objects via our databases.

Institutions often fear that free access to their digital holdings or even to the metadata on them might make them unnecessary and that people will no longer come to the reading rooms, sound archives, etc. I am convinced that the opposite is the case — using an institution via the Internet only means another type of user and usage. We cannot operate only locally, but regionally, nationally and internationally as well. This could cause the number of users to grow greatly — and these figures can be used for political argumentation and for generating political awareness.

But as a basic requirement for that we **must** provide free access to our catalogues and metadata.

In detail the metadata information includes:

- description of the content/provenance, etc. of the objects,
- description of the technical conditions of the object,
- description of the legal conditions under which the object can be used.

(For the rest of my speech I will concentrate on the first group of metadata.)

These metadata are essential because they are the connection between users on the one hand and objects on the other. The more information you can offer on the subject — on concrete content details, on the context of digital objects — the more people gain knowledge of the digital material in your institution. The wider you spread this information on your digital holdings, the more people like to have access to them, sometimes regardless of whether free or restricted access is possible. That means that your institution does not have to be the **one** single access point, but there should be others via related institutions, search engines and networks.

When preparing for this talk, I was asked whether librarians, archivists, curators of museums and other colleagues do not lose “control” of their digital holdings when giving free access to objects and data. And I was asked how to make sure that digital content and digital data are used in the right context. What do you think? Can we avoid usage of digital information that is different from our own professional or scientific thinking? Do we **have** to avoid this? I can imagine several forms of abuse of digital information from my point-of-view, but often others have different opinions.

And is the use of material in a different manner than usual not often the starting of new scientific discussions? How do you control such processes? No chance!

Access to digital content and data should be as open as possible in the light of copyright regulations and intellectual property rights. If not the objects themselves, the metadata should be freely accessible to make the content attractive to potential users.
Abuse cannot be prevented. Maybe we have some influence on the streams of data on the Internet to a certain extent, but no more.

Offering free access to metadata or even to digital objects themselves has also a lot to do with the question of “how best to prepare the search and retrieval of digital objects?” Traditionally search and retrieve is one of the main questions of our professions. We drafted cataloging or indexing rules over centuries, trying to make catalogues or finding aids as generally usable as possible. The description of objects was aimed at all necessary information our users needed to give them an idea of the single object or of the collection of which the object was a part. We still do this nowadays for analog material, and some institutions try to do the same for digital collections — but does that make sense?

For digitized material it is useful to use the metadata describing the originals. But the mass of born digital material is difficult to index or catalogue the same way. And even if you dare to be selective, it is much more material than ever before.

Additionally, you do not need to describe digital objects as you have done before. Users do not need to get an idea of the object, because, ideally, they can have direct access to it working in an archive or library, etc., or even via the Internet.

So far in my speech I have talked a lot about open access to our data, about Internet and networking, about users who are no longer local. With all that in mind, I believe that today's cataloging and indexing work consists of

- describing the necessary basic information
- analyzing the whole digital object and extracting automatically each additional piece of information you can get,
- combining this automatically generated information with the databases we already have (authority data, etc.),
- offering additional search entries by this.

We do not need to describe objects to include them in our catalogues or finding aids any longer; we need description to put the objects into subject or time or local contexts and offer as many access points from different points-of-view to it as possible.

Let me show you an example of the linking of data.

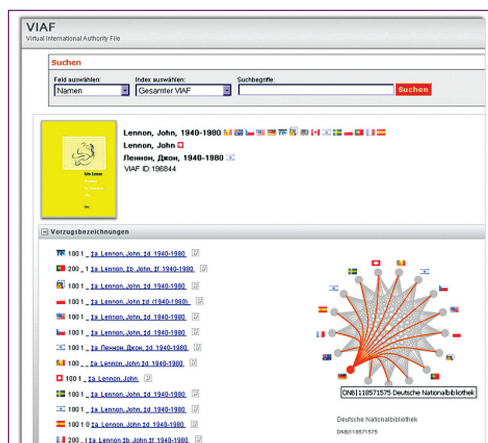


Figure 12: One of the international databases is the virtual international authority file where you then find information on John Lennon from other countries in the world.

In this sense metadata of originals that are digitized could be enriched by extracting additional information out of the content of digital files.

We at the German National Library actually are working on a project, PETRUS, that develops processes for automatic extraction or automatic setting of certain information. CONTENTUS contains extraction aspects for different purposes. Europeana or Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in the context of the German National Library are working on metadata enrichment by automatic processes. I do not want to concentrate on this now, but I'd like to point out that a lot of activities are undertaken in this field and the results can be re-used.

In any case, the metadata of digital objects have to include a persistent identification — that means a machine-readable, trusted address of the digital object that is kept alive “for eternity”. The situation of metadata containing an URL, behind which you see “Error 404 — file not found” is the worst in the digital world and should not happen to institutions like ours.

**Summarizing these aspects, I am convinced that sensible digitization also comprises the production of good metadata – intellectual or automatic. “Good metadata” in that sense means useful for our users, not for our catalogues.
To retrieve a digital file, you need persistent identifiers.**

After having done a lot of work on cataloguing, indexing and possibly enriching metadata, it is of high importance that the metadata can be found by users. For this, another decision in our institutions has to be made: the decision for an open, modern structure of our data formats and for allowing others — even so-called commercial search engines or WIKIPEDIA — to re-use our metadata.

Furthermore, it is necessary for the metadata to have an identifier too. And with this aspect of “re-use of data”, I bring in the idea of linked (open) data.

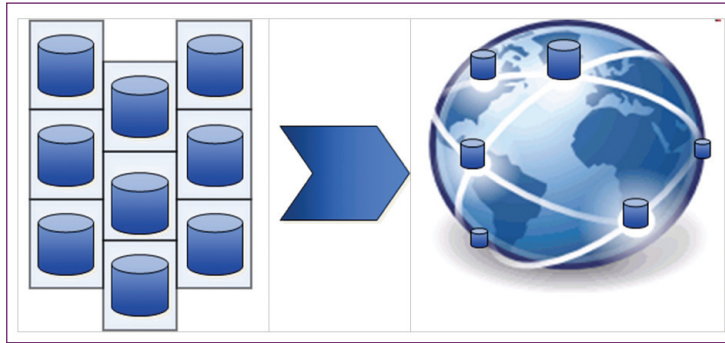


Figure 13: From single data silos to the Giant Global Graph (GGG) — the worldwide database of linked data

The idea of linked data is to move from closed parallel databases of single institutions to one worldwide database that consists of “networking”, of “linked data”.

In 2006 Tim Berners-Lee, the inventor of the Internet, stated: “*The Semantic Web isn’t just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.*”⁵

To realize this, he set up four rules:⁶

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Since then the development of the linked data community between 2007 and 2009 was as follows.

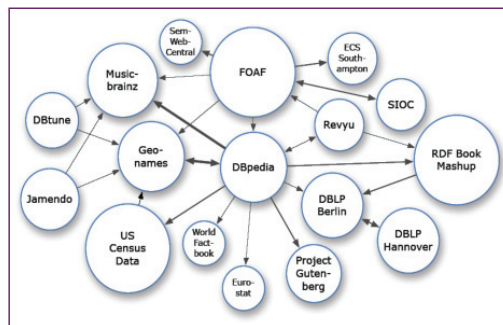


Figure 14: The community in July 2007

5 <http://www.w3.org/DesignIssues/LinkedData.html>

6 <http://www.w3.org/DesignIssues/LinkedData.html>

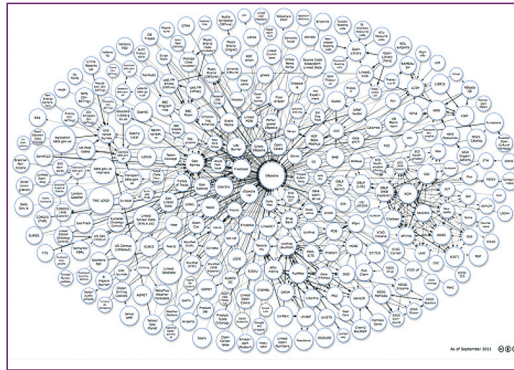


Figure 15: The community in September 2011

I am not the technical expert to explain this in detail, and this is not the right place to do such a precise presentation. But the ideas behind and the advantages of creating direct links between our data are:

- Standards for data publication and exchange already exist on the Web
- The possibilities of reaching more user groups increase (scientific institutions or communities, research institutions and communities, user groups behind other “linked data datasets” which are linked to your own)
- Your own data are enriched by information from other areas, e.g. geographical aspects, encyclopedias, thesauri, biographies, etc.

Practical examples for this can be found at the website of the BBC and at a new museum website in Germany. The BBC has a section about musicians and pop bands on its website, which shows the linking of data very well.⁷ The complete information is gathered from several sources, such as Wikipedia, Dbpedia, Musicbrainz and others.



Figure 16: BBC-Website on the Beatles

For example, if you look for “The Beatles” at the BBC-Website, you find

- the most actual up-to-date information concerning the group or the members of it (selected from the daily information of BBC),
- biographical details (Wikipedia),
- a list of tracks that have been played by the BBC during the past weeks (Musicbrainz),
- a list of the DJs who most currently played Beatles’ songs (Musicbrainz and daily broadcasting information)

⁷ <http://www.bbc.org>

- a buyers' guide
- a list of links to other information about the group
- etc.

Another example can be seen with the project “Museum-digital”⁸, which is currently in its beta version.

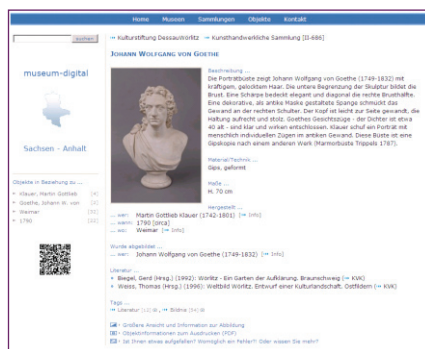


Figure 17: Information on a sculpture showing Goethe on the website of museum-digital

There you find additional information included in the object description on persons who are in any way at all involved with the objects. Or you get geographical details that come out of the “Getty Thesaurus of Geographic Names Online” (TGN). The connection between the description of the objects and the biographical information on persons is done by using the authority file for names of the German National Library.


I think these examples show very clearly the advantages of linking data from all over the world. It is not a linking of the objects themselves, but only of the metadata that describes the objects. Each bit of information contains the link to the objects or to a front page where the users can find a hint of how to achieve access.

This is a big service for users, and therefore I firmly believe that the re-use of data should not be limited to data exchange between certain databases. It should be developed to linked data activities – preferably ‘open’ – to build up the ‘Giant Global Graph’.

Well, my speech is based on my own opinion and the experiences we have had at the German National Library. You might have different point-of-views. But the summary of my recommendations on the sense of digitization is as follows:

Digitization makes sense, if

- as much material as possible is digitized,
- even objects which cannot be used in the original form, because of preservation aspects, are digitized,
- as much material as possible can be accessed via the Internet,
- the information (metadata) on the objects offers as many access points as possible,
- the metadata can be used by search engines and
- also re-used within the linked (open) data community.



Are we ready for that? Do we have the right curricula at our universities and professional schools? Do we address the right issues to our young professionals? Do we have to change the way we see ourselves within the different professions?

And, looking at the framework that is necessary to realize the picture I've drafted: Have we solved the legal or even the technical problems arising from such a picture? I am convinced that librarians, archivists, curators and other information specialists are going into the right direction, but nevertheless there is still a lot to be done.

John Lilly⁹ once stated: "Our only security is our ability to change." Our institutions have to change to play a further role in the changing information society.

The IASA 2011 conference in Frankfurt we just started will be a good platform to take one or more further steps, and I am convinced that we will hear a lot of new activities, learn a lot from the experiences gained from projects and discuss a lot about new developments and strategic directions.

⁹ John Lilly, http://en.wikipedia.org/wiki/John_C._Lilly