Towards Semantic Libraries

The Container, the Content and the Contenders

Prof. Dr. Stefan Gradmann
Humboldt-Universität zu Berlin / School of Library and Information Science
stefan.gradmann@ibi.hu-berlin.de
... à la carte:
A Three-Course Menu

- **Hors d'oeuvre**: Where are we coming from?
  Basics of Library Functionality
- **Main course**: Where are we going to?
  Semantic (Digital) Libraries
- **Dessert**: Who else is going there?
  Partners and Contenders
Hors d'Oeuvre

Where are we coming from?
Basics of Library Functionality
Library Functional Principles (1)

Metadata Catalogue

Document Objects

Author
Title
...ShelfNo

Author
Title
...URL

Author
Title
...URL

Author
Title
...IsShownAt

User

XML +
XSLT

PDF
Adobe

Europeana
Object

Towards Semantic Libraries / Prof. Dr. Stefan Gradmann
Tilburg: Digital Libraries à la Carte 2010
Library Functional Principles (2)

- **Mediating access** to information objects via **catalogues**
- **Mediating links** as pointers from metadata to objects
- Objects are part of a library **collection**
  - An object to be used within a library typically is part of this library's collection
- Internal processing logic: focus on
  - objects as information **containers**,
  - not so much on the **content** of these containers
- **Ingestion, storage, description** and **retrieval** of information objects as functional macro-primitives
Main Course
Where are we going to? Towards the Semantic Library
Linear Document Continuum ... … in the Gutenberg galaxy
Linear Document Continuum ...

... in emulation mode

- 'write' (Office / LaTeX)
- read + 'write' (Office / LaTeX)
- 'print' (PDF)
- 'write' (Office / LaTeX)
- 'write' (Office / LaTeX)
- quote
- publish
- read
- manage / organise
- library automation
- review
- apprehend
Linear Document Continuum ... … going digital (entering Turing galaxy)

e-annotate - as part of document? Linked to it? How?

replicate or 'point to'? How to address microstructures??

'read' ??

annotate

quote

apprehend

XML + XSLT

generate XML / XSLT

author

review

publish

manage / organise

digital libraries??

stabilise, version, add identifier

e-annotate (public?)

Beyond and far beyond / Prof. Dr. Stefan Gradmann
Tilburg: Digital Libraries à la Carte 2010
Decreasing functional determination by traditional cultural techniques
Disintegration of the linear / circular functional paradigm
Erosion of the monolithic document notion in hypertext paradigms
Ted Nelson's Xanadu: radicalised Hypertext ...
... the web of 'documents' extended with a web of 'things' ...

http://dbpedia.org/page/Cat

"cat"

http://upload.wikimedia.org/wikipedia/commons/5/58/Collage_of_Six_Cats-1.jpg

dcterms:description

HTTP 303 / Accept: application/rdf+xml

http://myspace.com/RealCats

HTTP 303 / Other content types or no Accept: Header
... and 'publication' aggregations combining 'documents' and 'things'.

Where do resource aggregations 'start'? Where do they 'end'?

And what constitutes document boundaries??

And which node was connected to which one at a given time???
Machines can reason on triple sets!

http://ex.org/bib/Book

rdf:type

http://ex.org/bib/White_Noise
Some reasoning preconditions ...

http://ex.org/bib/Book

http://ex.org/bib/White_Noise

http://ex.org/bib/ArtisticWork

rdf:type

rdfs:subClassOf
... and an automated inference!

There is quite some potential for generating scholarly heuristics here!
The use of Inferences

LoD: Billions of Triples …

… and Semantic Publishing!

http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets
Semantic Publishing as Defined by Shotton

Shotton et al. (2009b) define semantic publication to include anything that

- enhances the meaning of a published journal article,
- facilitates its automated discovery,
- enables its linking to semantically related articles,
- provides access to data within the article in actionable form, or
- facilitates integration of data between articles.

Example of an enhanced article
Behind the Scenes

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slums

*Shotton et al.* 2009, Adventures in semantic publishing

*Reis et al.* 2008, original

Data fusion

Google Maps mashups

PLOS COMPUTATIONAL BIOLOGY

SEMANTICALLY ENHANCED VERSION OF A RESEARCH ARTICLE FROM PLOS NEGLECTED TROPICAL DISEASES

Table 1 data

Fig. 2 data

Fig. S2 data

S1: Technical details

S2: Annotation guidelines

RDF1: Details of article

RDF2: Details of citations

Downloadable datasets

Interactive Fig. 3

Document summary
Semantic Enrichment Tools

- **Generic:**
  - OpenCalais (http://www.opencalais.com/)
  - Temis (http://www.temis.com/)

- **Specialised:**
  - ConceptWebAlliance (http://conceptwiki.org) (Biomedical, Jan Velterop)

- Good critique by Roderic Page:
  - “linking terms to HTML pages doesn't get us much further. Great for humans, not so good for computers.”
  - Too much focus on journal article format!
  - → We need a little more!
“Turning inked letters into electronic dots that can be read on a screen is simply the first essential step in creating this new library. The **real magic** will come in the second act, as each word in each book is
- cross-linked,
- clustered,
- cited, extracted,
- indexed,
- analyzed,
- annotated,
- remixed,
- reassembled

and woven deeper into the culture than ever before. In the new world of books, every bit informs another; every page reads all the other pages.”

A **Semantic Wiki**,  
not based on static HTML pages, but instead consisting of **dynamic documents**,  
provided at runtime from **semantic microcontent** (“PAUX-Objects”),  
**semantically linked** by “PAUX-Links”  
Microcontent elements have **HTTP URIs**!  
→ PAUX documents can be published as Linked (Open) Data aggregations with maximum granularity: down to **word level**.  
**PAUX creates “liquid books”**  
PAUX (http://www.paux.de): origins in eLearning
Granular Semantic Publishing: Paux (1)
Very Granular Semantic Publishing: Paux (2)
Semantic Publishing: Paux (3)
Linked Semantic Publishing: Paux (4)
Linked Semantic Publishing: Paux (5)
Social Semantic Publishing: Paux (6)
Towards Semantic Libraries / Prof. Dr. Stefan Gradmann
Tilburg: Digital Libraries à la Carte 2010

Paux live (1): Outline & Sentences

PAUX is an innovative Content Management System that stores text modularly.

PAUX is a semantic and distributed but centralized content by linking reusable semantic content objects semantically.

PAUX takes a own path of knowledge representation by linking single objects such as words, sentences, pictures, persons etc. to a network.

Therefore, PAUX can be classified only partially in the area of existing systems like CMS/ECMS, Wiki, LMS or LCMS. The semantic content objects of PAUX are to make knowledge available as filterable content for Websites, semantic Wiki, detailed-evaluated eLearning and individualized print media.

PAUX is a knowledge management system written in Java and addresses enterprises (also publishing houses) as well as colleges, offices and other authorities.

It is free of charge in the context of a development partnership.
Paux live (2): Sentence & Linking Options
Paux live (3): Word & Hyperlinks
Paux live (4): Word & Link to Sentence

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Wort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PAUX</td>
</tr>
<tr>
<td>2</td>
<td>is</td>
</tr>
<tr>
<td>3</td>
<td>an innovative Content Management System that stores text modularly.</td>
</tr>
</tbody>
</table>

This is made possible by the fact that PAUX divides text content into its smallest components (up to word-level) and links them semantically.
Data = Publication

- Distinction data vs. publication will get increasingly obsolete in semantic publishing environments …
- … at least in the STM sector.
- The move into semantic publication will be much slower in the SSH because of:
  - fuzzy and unstable terminology
  - fuzzy linking semantics hard to formalise consistently
  - close relation between complex document formats and scholarly discourse
- Current examples are mostly from the medical and biomedical area as a consequence.
- => Jan Velterop's concept of “Nano-Publications” or Bill Town's examples from chemistry (namely OreChem)
“What do you do with a million books?” (G. Crane)

- DL view: digitisation and (more and more) semantic publishing increase by at least one or even more orders of magnitude
  - Scale
  - Linguistic heterogeneity of content
  - Granularity of objects
  - Noise (encoding and semantic)
  - Audience
- They may lead to a dramatic decrease of the number of collections and distributors
- They render obsolete the very notion of a 'collection' ...
- … as well as the notion of a 'catalogue'
- → Do we need more than one Digital Library in such a setting?
Re “What do you do with a million books?” (G. Crane)

- Scholarly view: digitisation and (increasingly) semantic publishing result in
  - growing quantity
  - increased complexity
- Well beyond scholarly processing capacity (=reading faculty)
- Multiplication of collections or distributors is annoying → as few as possible. Ideally just one (?)
- Scientists and Scolars will badly need help in these areas:
  - **Semantic abstracting, named entity recognition** for “strategic reading” (Renear)
  - **Contextualisation** of information objects
  - Robust **reasoning and inferencing** yielding digital heuristics

=> **Opportunities for libraries … but for others, too!**
Dessert
Who else is going there?
Partners and Contenders
Partners and Contenders: A Mostly Ambivalent Scenario

- Commercial enterprises as **partners**
  - Document mining, named entity recognition and semantic aggregation (OpenCalais, Temis and others)
  - Search engine technology (Google and others)
  - Library automation suppliers (OCLC, ExLibris and others)

- Commercial enterprises as **contenders**
  - Google
  - Amazon
  - Publishers
  - OCLC

- Will other digital libraries be partners or contenders?
Conclusion:
Towards Semantic Libraries

Another triple paradigm shift

- Focus on **container** → focus on **content**
- Metadata **catalogue** → **semantic network** of contextualized object representations
- Information object **storage and retrieval** → **knowledge generation**

Or else the WWW will share the fate of the Concorde aircraft or of the Zeppelin (Eco) and **nothing of this is going to happen** …

… I don't believe so!
Suggested Reading


Thank you for your patience and attention