Building Blocks of the future Scholarly Web: Beyond and far beyond

The APE Lecture 2010

Prof. Dr. Stefan Gradmann
Humboldt-Universität zu Berlin / School of Library and Information Science
stefan.gradmann@ibi.hu-berlin.de
What's On the Menu?

- **Hors d'oeuvre:** Evolution and de-construction of the printed document continuum – from the Gutenberg Galaxy into the Turing Galaxy and further beyond into Semantic Publishing
- **Main course:** Le document selon Roger – the work of RTP-Doc on document deconstruction
- **Dessert:** Why we need the humanities to understand all this … and why understanding this process is vital for the humanities!
Linear Document Continuum ... … in the Gutenberg galaxy
Linear Document Continuum ... 
... in emulation mode
Linear Document Continuum ... 
... going digital (entering Turing galaxy)

- e-annotate - as part of document? Linked to it? How?
- generate XML / XSLT
- review (public?)
- publish
- stabilise, version, add identifier
- digital libraries??
- 'read' ??
- apprehend
- quote
- replicate or 'point to'? How to address microstructures??
- annotate

XML + XSLT
Web Based Scholarly Working Continuum ... 
... a triple paradigm shift: moving beyond

- Decreasing functional determination by traditional cultural techniques
- Disintegration of the linear / circular functional paradigm
- Erosion of the monolithic document notion in hypertext paradigms
'Document' Disintegration: moving far beyond towards RDF

- Further erosion even beyond hypertext: 'Documents' vs. RDF triple sets

Diagram:
- http://ex.org/bib/Book
- rdf:type
- http://ex.org/bib/White_Noise

Beyond and far beyond / Prof. Dr. Stefan Gradmann
Berlin: APE 2010
Some basic inferencing (1)

- http://ex.org/bib/Book
  - rdf:type
    - http://ex.org/bib/White_Noise
  - rdfs:subClassOf
    - http://ex.org/bib/ArtisticWork
Some basic inferencing (2)

http://ex.org/bib/Book

http://ex.org/bib/White_Noise

http://ex.org/bib/ArtisticWork

rdf:type

rdfs:subClassOf

rdf:type
LoD: Billions of Triples … … and Semantic Publishing!
E. g. OAI-ORE Resource Aggregations: what constitutes document boundaries?

- 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???
How this paradigm shift affects science and scholarship

- Four reasons for the erosion of the monolithic 'document' notion in hypertext and linked data paradigms:
  - **Distribution** of document resources in network based environments;
  - loss of the constitutive **linearity** almost impossible to compensate;
  - loss of the **unity of medium and content** (which – together with linearity - was a guarantee of **document integrity** in the Gutenberg galaxy)
  - Atomization of complex document entities in RDF triple environments.

- The erosion of the document notion in a digital network based context will affect science and scholarship differently:
  - It doesn't impact science a lot: document and signification models always have been a secondary concern in an area that currently is re-focussing on publishing data together with processing methods rather than 'documents'
  - It fundamentally impacts the humanities in that it changes the **conditions of apprehension and reuse** as well as the **basic modes of signification** of documents as complex signs.

- This process of erosion and the conditions required for reconstructing the document notion in a digital, network based setting were the main concerns of R. T. Pédauque.

Philosophy, linguistics, semiotics, computer science, mathematics, library and information science.

Evolution of the 'document' notion in the passage from printed to digital to web documents along three not mutually exclusive paradigms

- **Form** (vu='Look at', morphosyntax), as material or non-material structured object;
- **Sign** (lu='read', semantics), as meaningful instance and thus both intentionally determined and part of a sign system;
- **Medium** (su='Knowledge, Interpretation, Apprehension', Pragmatics) as a vector of communication, part of a social reality with constituting temporal and spatial processes of mediation.
Equation Triplet 1: Document = Form

- **Document = Form**: object nature is constitutive
  - Traditional document = medium + inscription
  - Electronic document = structures + data
  - XML-document = structured data + stylesheet
  - Definition: *An electronic document is a data set organized in a stable structure associated with formatting rules to allow it to be read both by its designer and its readers.*

- **Agenda**
  - To what degree can we conceive structures and data independently from each other? How strongly do we need to relate them to each other to create digital 'readability'?
  - How device-independent is 'readability' (W3C fights about XHTML!)?
  - Long term preservation: structure and/or data and/or styles?
  - How far can we push from precoordination to postcoordination?
  - To what extent is the analogy with established cultural object formats costitutive for human perception (e. g. PDF)?
Equation Triplet 2: Document = Sign

- **Document = Sign**: meaningful nature is constitutive
  - Traditional document = inscription + meaning
  - Electronic document = informed text + knowledge
  - Semantic Web document = informed text + ontologies
  - Definition: *An electronic document is a text whose elements can potentially be analysed by a knowledge system in view of its exploitation by a competent reader.*

- **Agenda:**
  - 'Document' becomes a secondary notion: the primary interest is with 'text' and its 'context', which together enable interpretation.
  - Relation of text and markup, of text and metadata.
  - Meaning and reference / denotation: a critique of semantic web signification models.
  - How to generate trust and authenticity?
Document = Medium: document als social phenomenon

- Traditional document = inscription + legitimacy
- Electronic document = text + procedure
- Web-Document = publication + measured usage/access
- Definition: *An electronic document is a trace of social relations reconstructed by computer systems.*

Agenda

- Role of archives in this context: document the past or record things happening?
- What sense do the borders between archives and publication still make?
- How to generate 'attention'?
- Who pays? And what for?
“A document may finally be nothing more than a contract between people. This contract has
- anthropological (legibility-perception),
- intellectual (understanding-assimilation) and
- social (sociability-integration)
properties which may form the basis for part of their humanity, their capability to live together.“

RTP Doc is concerned with the **de-construction of the document notion** in digital, distributed settings.

The text also challenges the **monolithic notion of a personal author**: behind the pseudonym Roger T. Pédauche more than 60 individual researchers have been actively involved in the drafting of the three volumes.
Conclusion: Document De-construction, Building Blocks - and the Humanities

- RTP Doc provides rich lines of thought that may ultimately enable us to reconstitute a meaningful notion of 'documents'
  - as complex but discrete entities made up of aggregated digital resources
  - including deconstructed notions of integrity and authenticity
  - but which do not borrow on the print analogy anymore.

More generally, this means reformulating the terms and conditions of the readability contract – and this in turn is vital for the humanities disciplines grounded in interpretation!

- Without consistent and clearly established notions of signification and interpretation relating to well understood document objects these disciplines cease to exist.

- Without these humanities disciplines it will be impossible to re-establish a meaningful document notion in digital environments!

- Lire Roger!

- Move well beyond – at least regarding RDF and Semantic Publishing

- And thank you for your patience and attention.