

Simple Semantic Enrichment of Scientific Papers in Social Sciences

[Alexander Garcia](#) / [Philipp Mayr](#) / [Leyla Jael Garcia](#)

Florida State University / GESIS / biotea.ws



Outline

- * Motivation
 - * What data do we have?
 - * Why we are doing this?
 - * What are we doing? What do we aim to achieve?
- * RDF generation
 - * Metadata and Content
 - * Content enrichment
- * Consuming and delivering the data
 - * A first approach

Motivation

What data do we have?

- * GESIS
 - * Leibniz Institute for the Social Sciences
 - * Support for the research cycle
 - * Journals: ISI, MDA
- * MDA – Methods, Data, Analysis
 - * Journal for Empirical Social Science Research
 - * Focus on
 - * Survey methodologies
 - * Methods in empirical social research
 - * Open-access, full-text

Motivation

Why we are doing this?

- * The World Wide Web
 - * Dissemination infrastructure: Scientific and non-scientific contributions
 - * Information:
 - * Still locked up in discrete documents
 - * Not interconnected, not machine-processable
- * RDF technology:
 - * Connectivity tissue
 - * But how does it impact to the scientific communication?

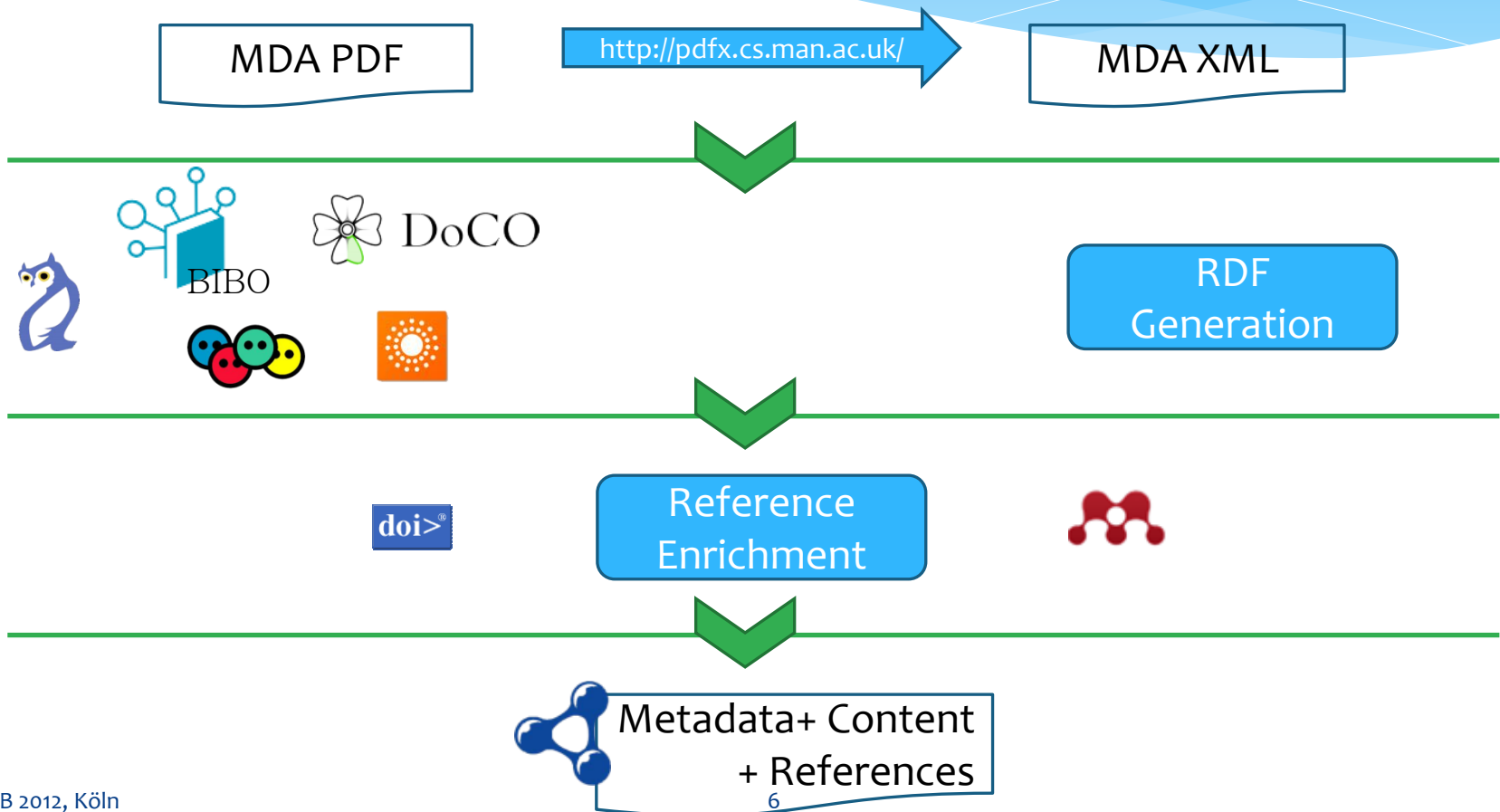
Motivation

What are we doing? What do we aim to achieve?

- * Question: How can scientific publications be delivered into the Semantic Web?
- * Our approach
 - * RDF for research articles
 - * Entry point to the Web of Data
 - * Part of the Linked Open Data
 - * Semantic enrichment
 - * Interoperable with online data
 - * Richer user interface
 - * A different read experience
 - * Interconnected with external related elements
 - * Collaborative environment

RDF Generation

Metadata and Content



RDF Generation

Content enrichment

 Metadata+ Content
+ References



Automatic
Annotation



Manual
Annotation



 Automatically Annotated RDF

 Manually Annotated RDF

Lessons learnt

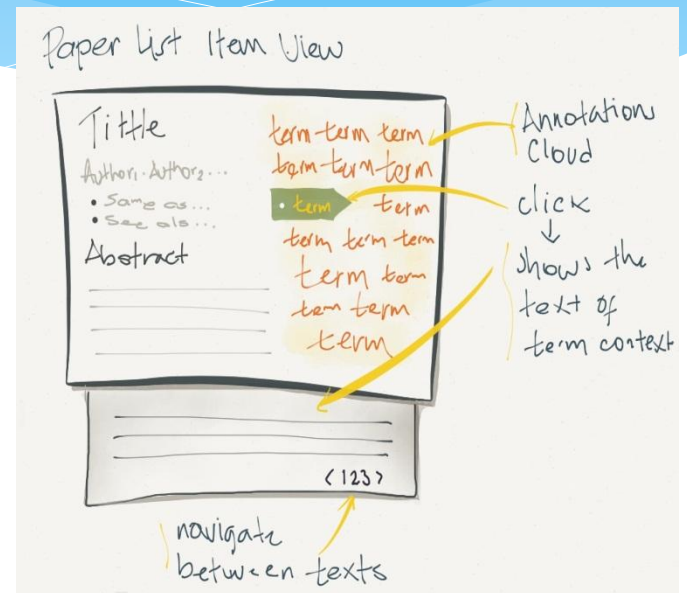
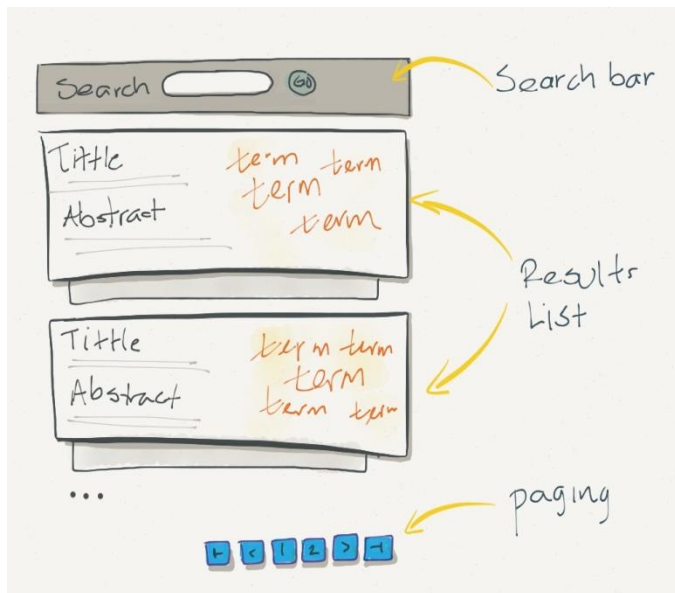
- * Biotea, a similar project on the biomedical domain
 - * XML to RDF works well
 - * RDF annotation works well but ... annotators are not perfect
 - * Format is not translated → bold, italics
 - * Modeling tables is not easy
 - * Dictionary-based entity recognition tools works better
- * This project
 - * PDF to XML is not perfect

Consuming and delivering the data

- * What does it make possible ?
 - * How similar are two articles? → based on concepts → semantic similarity
 - * What articles use this reference in a section with title “Results”?
 - * Which annotation co-occurs more with this “X” annotation?
 - * Which articles include term “A” but not term “B”?

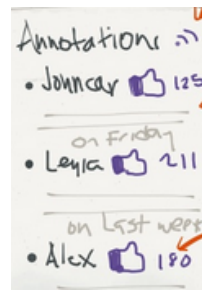
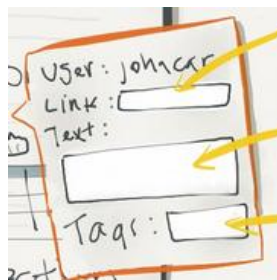
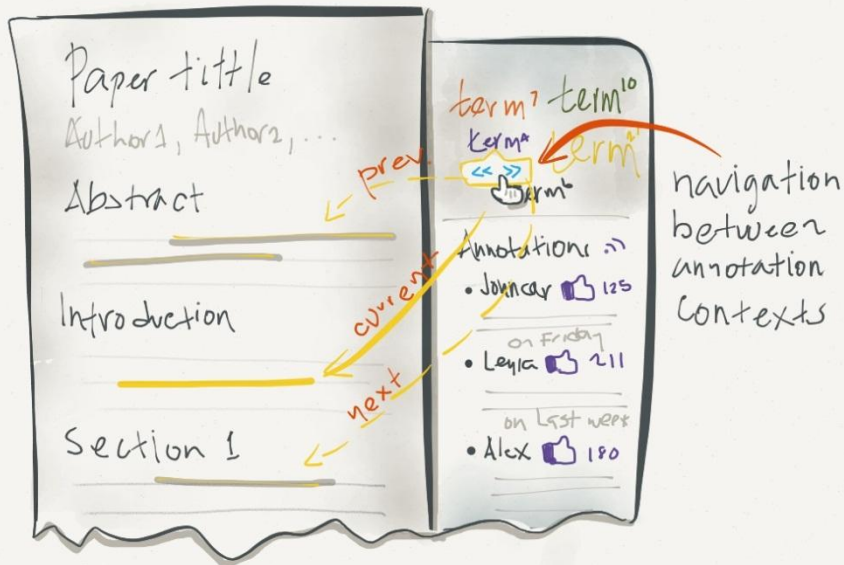
Consuming and delivering the data

A first approach



Consuming and delivering the data

A first approach





- * Contact

- * Alex García, alexgarciac@gmail.com
- * Philipp Mayr, philipp.mayr@gesis.org