Digital Archives Supporting Document Content Inference

Evgeny Cherkashin, Alexey Shigarov,
Viacheslav Paramonov, Andrey Mikhailov

V.M. Matrosov's Institute of System Dynamics and Control Theory SB RAS

CIS, MIPRO-42, 20-24 May 2019, Opatija, Croatia

Document authoring and storage

In most cases documents are created as a result of

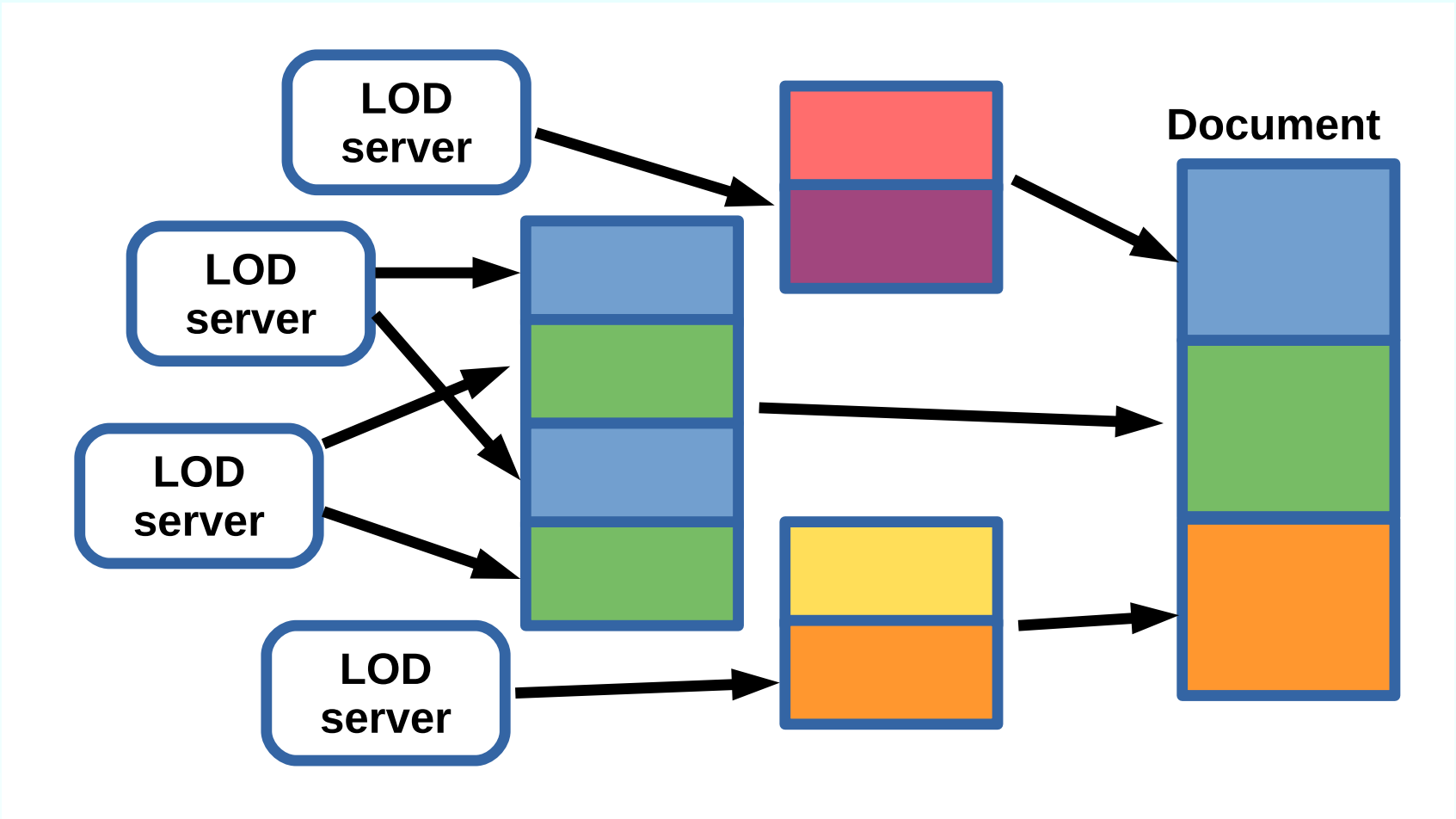
- creative activity of a person with a text processors (authoring);
- printing a digital copy or a data record in a database;
- aggregation operation over database records (report).

Then it is stored either as a physical paper and/or a digital document (PDF, DOCX, HTML).
Since 2000-th, Semantic Web and Linked Open Data (LOD) is being developed, allowing

- structural storage of data within published documents;
- processing stored data computationally;
- integration of data structures and data objects globally.

The ***aim of this research*** is to develop technologies, software and services allowing construction of digital archives supporting document data inclusion and inference from existing documents.

Structure of a document



Linked Open Data

1. Information is published in Internet with open access license;
2. It is represented in a machine-readable form, e.g., Excel table instead of a bitmap picture;
3. An open format used, e.g., CSV instead of Excel;
4. The format is based on W3C recommended standards, allowing RDF and SPARQL reference;
5. Published data refer to objects, forming context.

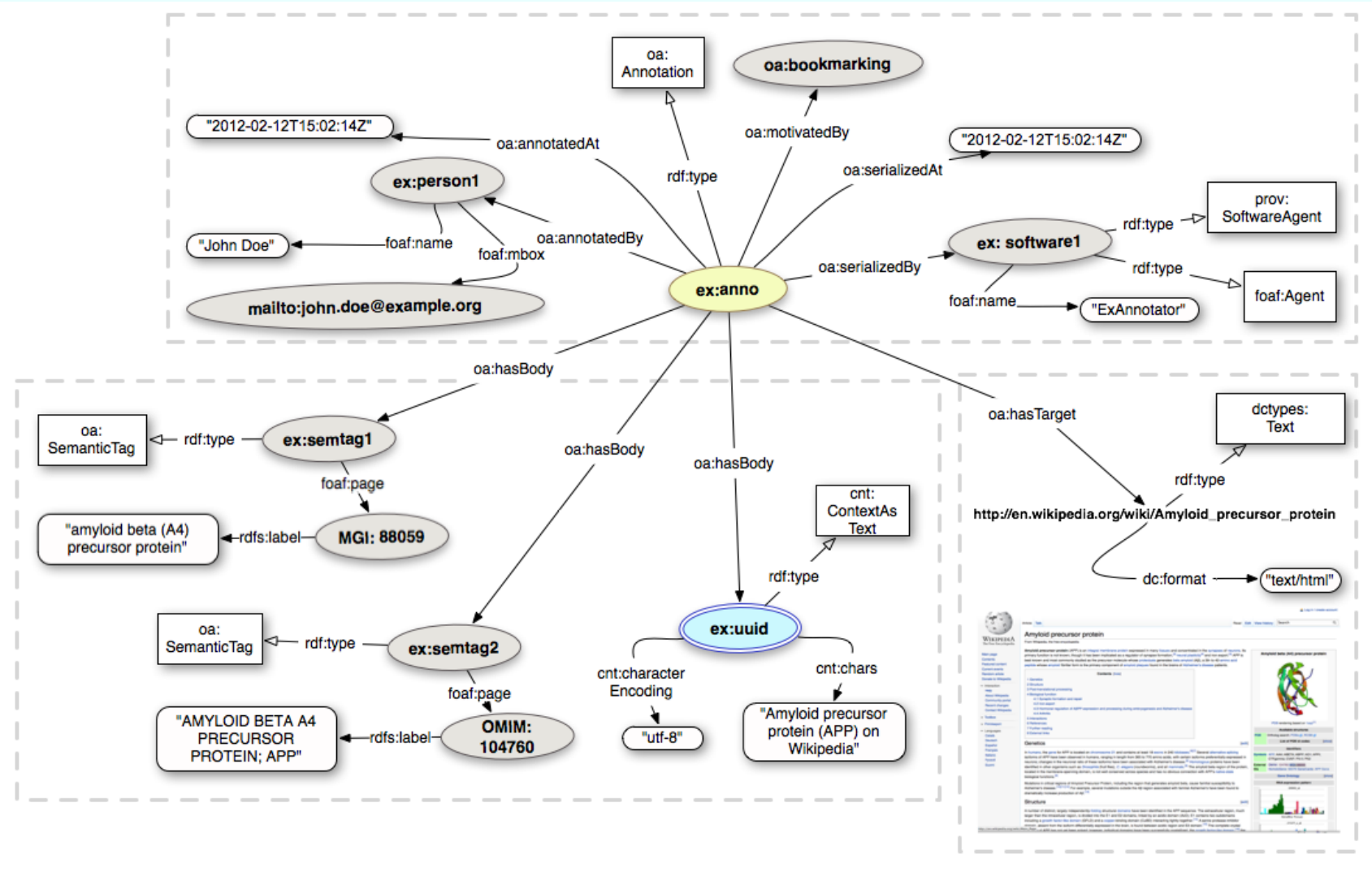
Thus, applications publish data as relations of objects (entities)

Contribution

Our work presents a design of digital archives, which allows developers device information system and document processing services with the following features:

- load LOD marked up document, extract, store in a graph and index RDF data;
- retrieve RDF data as triples or as a result of full-text search query;
- combine existing LOD data and its content in new documents dynamically thanks to relatively simple browser based context inference machine;
- ability to use server site inference machine (Prolog) to process RDF data upon request from browser's part of the system;
- convert created RDFa marked up HTML5 documents into Excel and Word formats.

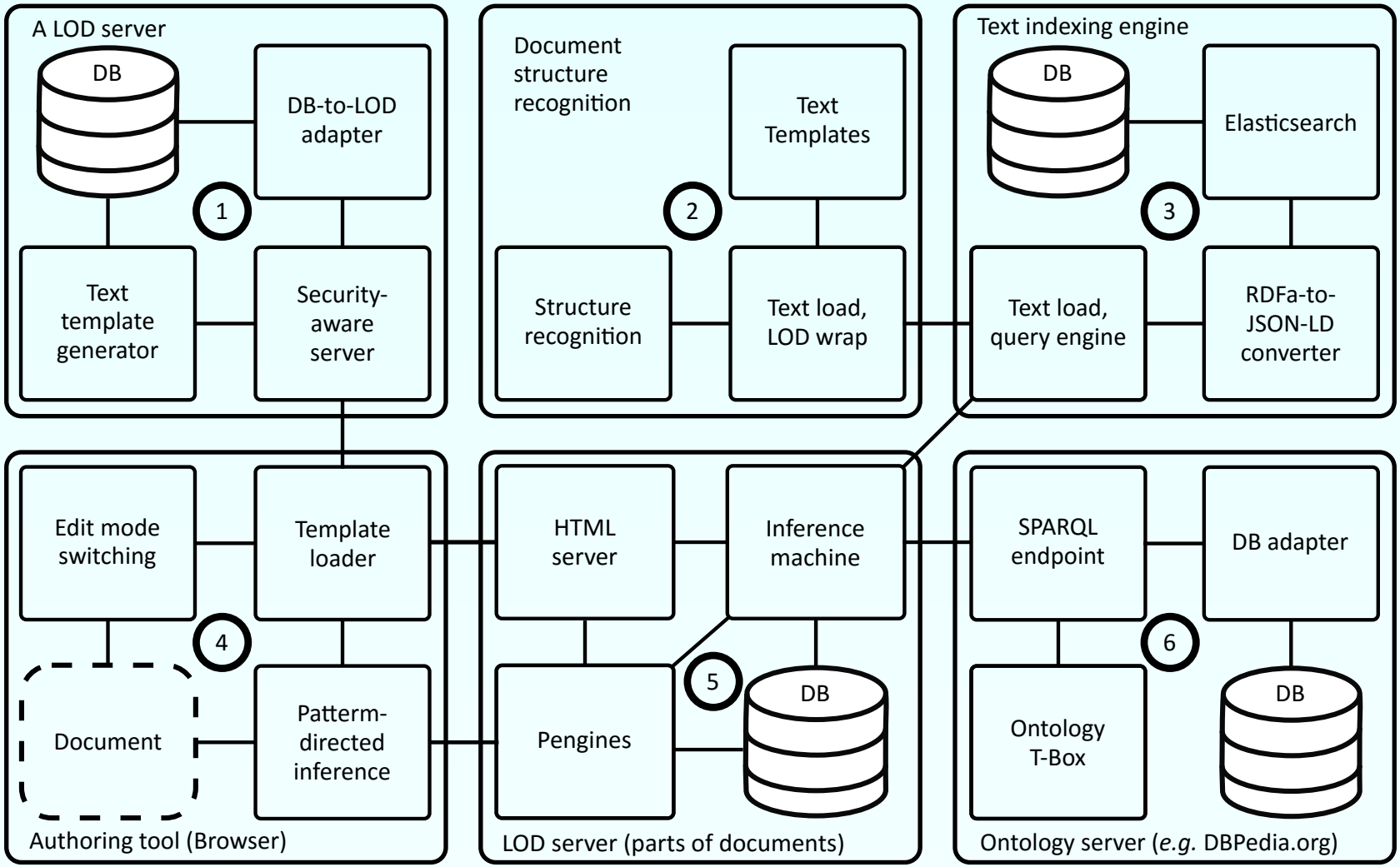
Open Annotaiton (oa)



Representation

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 ...>
<html lang="ru" xmlns=http://www.w3.org/1999/xhtml
xmlns:taa=http://irnok.net/engine/rdfa-manipulation
xml:lang="ru" metal:define-macro="page">
<head>
<!- Connecting stylesheets and modules ->
</head>
<body prefix="rdf: http://www.w3.org/1999/...-ns#
foaf: http://xmlns.com/foaf/0.1/ imei: imei.html#
course: https://irnok.net/college/plan/01..16-...\
%D0\%BA_PB-SM.plm.xml.xlsx-....2.3.1.html#"
resource="#post"
typeof="schema:CreativeWork sioc:Post prov:Entity">
<!- The application control panel ->
<main lang="ru" resource="#annotation"
typeof="oa:Annotation" id="main-doc-container">
<div property="oa:hasTarget" resource="#course-
work-prog"></div> <article property="oa:hasBody"
typeof="foaf:Document curr:WorkingProgram"
resource="#course-work-program" id="main-document">
<div taa:content="imei:title-page"></div>
<div taa:content="imei:neg-UMK"></div>
<section id="TOC" class="break-after">
<h2 class="nocount c">Table of Contents</h2>
<div id="tableOfContents"></div>
</section>
<section id="course-description"
  resource="#description"
  property="schema:hasPart"
  typeof="schema:CreativeWork">
<div property="schema:hasPart" resource="#purpose"
  typeof="dc:Text cnt:ContentAsText" >
<div property="cnt:chars"
  datatype="xsd:string">
<h2 property="dc:title"
  datatype="xsd:string">Aims and objectives of the
discipline (module)</h2>
<p>The aim of teaching the discipline ...</p>
</div> </div>
. . . . . .
```

Architecture



Examples of documents1

Examples of documents2

Examples of documents3

Conclusion

Thanks for Your attention!