Integrating external sources in a corporate semantic web managed by a multi-agent system

Tuan-Dung CAO[1], **Fabien GANDON**[1,2]
[1] ACACIA Team, INRIA Sophia Antipolis
[2] Computer School, Carnegie Mellon University

Fabien.Gandon@sophia.inria.fr

Introduction

Research problems in ACACIA team:

- Organisations **need** to adapt to an ever changing world
  - **Nervous system**: capture and diffuse knowledge
  - **Persistent memory**: store and/or index knowledge
- Study problematics of **organisational memories**
- Here: special case of external **organizational knowledge**

You are here

Plan

Brief summary of previous work on **CoMMA**

- **Corporate semantic Webs** as corporate memories
- Multi-agent system as **management architecture**

New society of **wrappers / HTML scrappers**

- **XML-based extraction** process
- Wrapper society: roles and interactions

- Dynamically **integrating heterogeneous sources** of information
  OBSERVER [Mena *et al.*, 1996] InfoSleuth [Nodine *et al.*, 1999] Carnot [Collet *et al.*, 1991] InfoMaster [Genesereth *et al.*, 1997] SIMS [Arens *et al.*, 1996] RETSINA [Decker & Sycara, 1997] Manifold [Kirk *et al.*,1995]

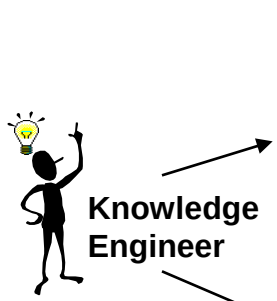
- Assist the management of digital **libraries**
  SAIRE [Odubiyi *et al.*, 1997] UMDL [Weinstein *et al.*, 1999]

- **Organisational knowledge** management:
  - **Collaborative** gathering, filtering and profiling
    CASMIR [Berney & Ferneley, 1999] Ricochet [Bothorel & Thomas, 1999]
  - **Mobile** access & domain model for **document classification**
    KnowWeb [Dzbor *et al.*, 2000]
  - **Taxonomy** of topics, profiling and **push** RICA [Aguirre *et al.*, 2000]
  - **Ontology** and **corporate memory**:
    multiple ontologies FRODO [Van Elst & Abecker, 2001]
    semantic intraweb, ontology, user profiling

CoMMA

Implementation choices:

- **Materialisation** of memory RDF(S) and its XML syntax (manipulated with **CORESE**)



RDF Schema → Support in CGs

RDF Rules

RDF Annotation

principle

example of schema

```
<rdfs:Class rdf:ID='Entity'/>
<rdfs:Class rdf:ID='Document'>
  <rdfs:subC
</rdfs:Class
<rdf:Propert
 <rdfs:subPr
  <rdfs:doma
  <rdfs:rang
</rdf:Proper
```

example of rule

```
IF
   Group
    Include
     Person ?x
    Include
     Person ?y
```

```
ity' />

reator' />
t'/>
>
```

example of query annotation

```
<CoMMA:Document rdf:about=
 <CoMMA:Title>Annual activ
 <CoMMA:Author>
  <CoMMA:Person rdf:about=
 </CoMMA:Author>
</CoMMA:Document>
```

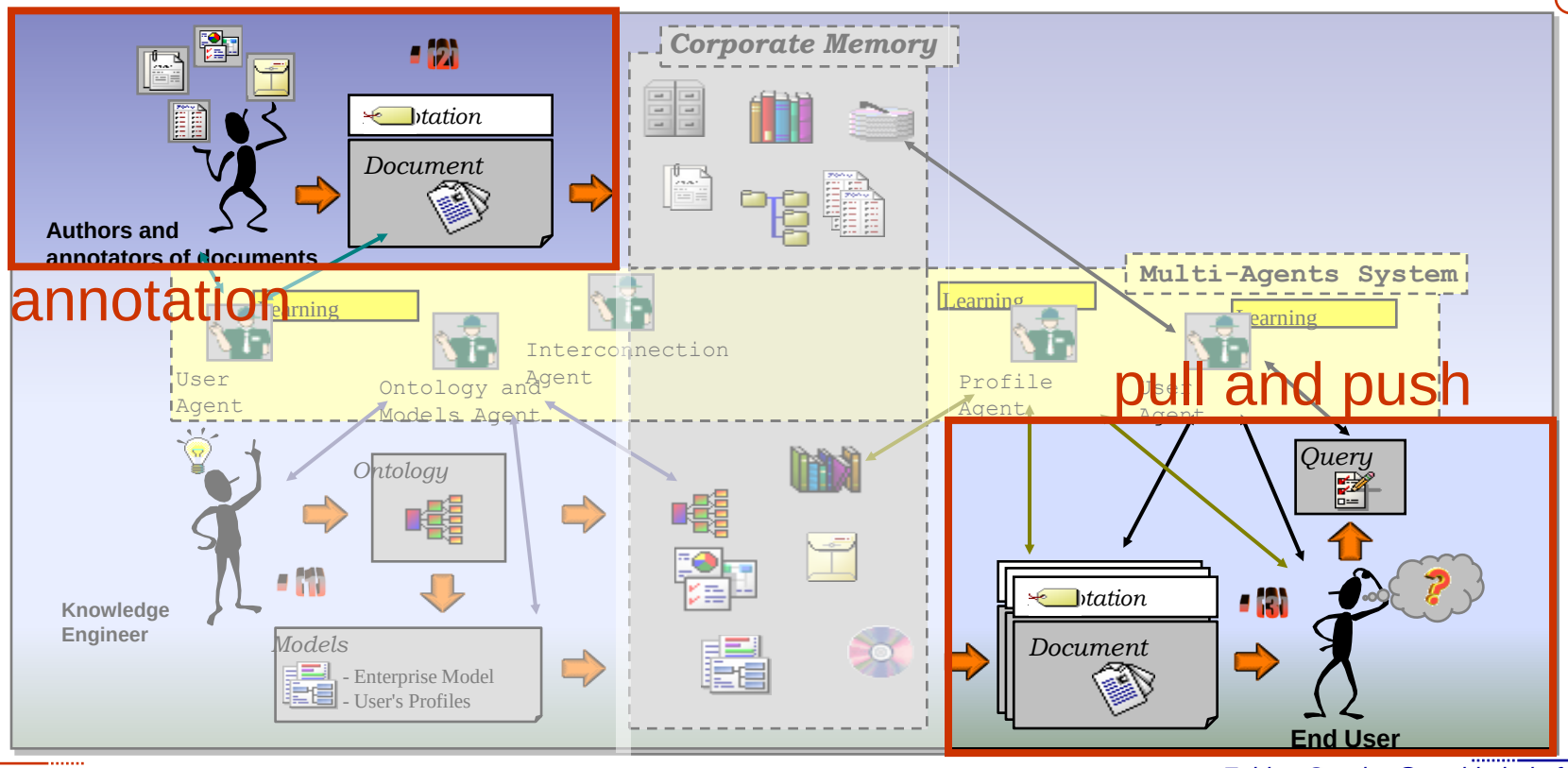
```
<CoMMA:Article rdf:about="?d">
 <CoMMA:Title>?t</CoMMA:Title>
 <CoMMA:Author>?a</CoMMA:Author>
 <CoMMA:Concern>
  <CoMMA:UMTSTopic />
 </CoMMA:Concern>
</CoMMA:Article>
```

Translation Projection Translation

Implementation choices:

- **Materialisation** of memory
  RDF(S) and its XML syntax (manipulated with **CORESE**)
- **Exploitation** of memory (DAI Archi. adequate D. Memory)
  Multi-agent system and machine learning techniques
  (implemented with JADE fipa) (implemented with **WEKA**)

overview



Ontology & model society: **replication**

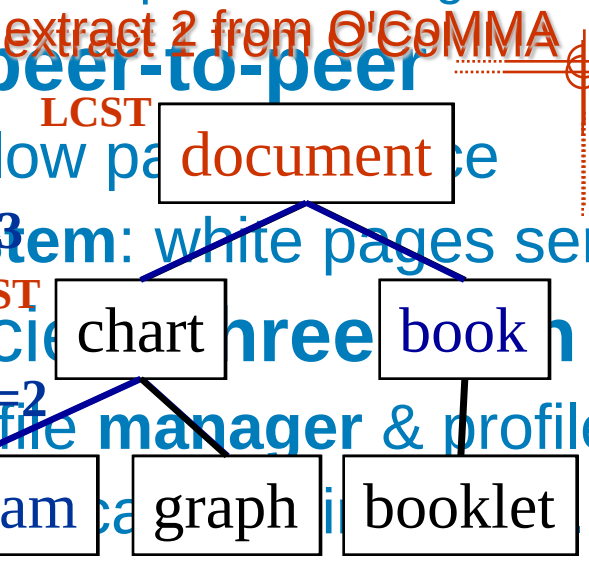
- **Ontologist**: to **store** and **provide** ontology
- **Corporate model archivist**: to **store** and **provide** structural model of the (human) organisation

Annotations society: **hierarchy**

- **Archivist**: handles local annotation sources / archives
- **Mediator** manage distributed processes for 2 tasks:
  - new annotation submissions (contract-net and semantic distance)
  - query solving processes (decomposition using URI as cut/joint points)

Matchmakers society: **peer-to-peer**

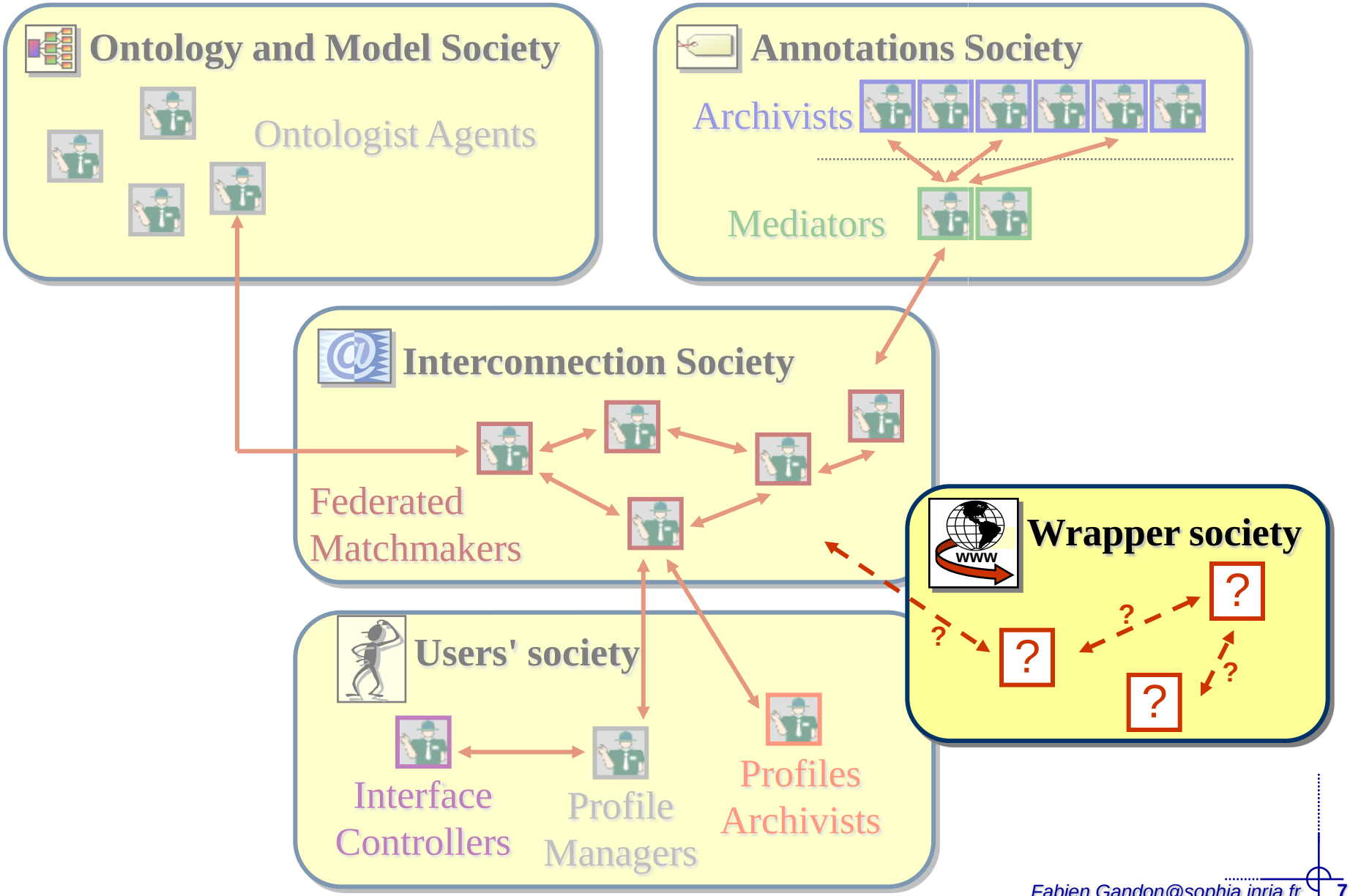
- **Directory facilitator**: yellow pa ce
- **Agent management system**: white pages service



User-dedicated sub-soci hree roles

- Handle users' profile: profile **manager** & profile **archivist**
- **Interface controlle** ca i

Roles and interactions



- No organization is an island: in a society, a market…
- Information resources on the **open Web** relevant
- **Integrate** external resources in memory *i.e.* annotate
  - Corporate portal: outer → inner *vs.* **inner → outer**
  - CoMMA: scenarios with **annotation roles** (e.g., librarians)
  - Large sets of documents ⇒ annotate = tedious repetitive task
    Manual annotation unrealistic
  - Repositories, digital libraries, *etc.* : recurrent **structural** clues
  - Annotations mainly based on information present in the page
  - Automate some **extractions** in **rules** to scale-up annotation
  - For pages of one site: "**annotate one, extract every others**"

  ⇒ Introduction of a new society of wrappers / scrappers

<InAnAsideToTheReviewers>

Multiple ontology & mapping: another **on-going work**

Focus on **one use scenario** with **one appropriate ontology**

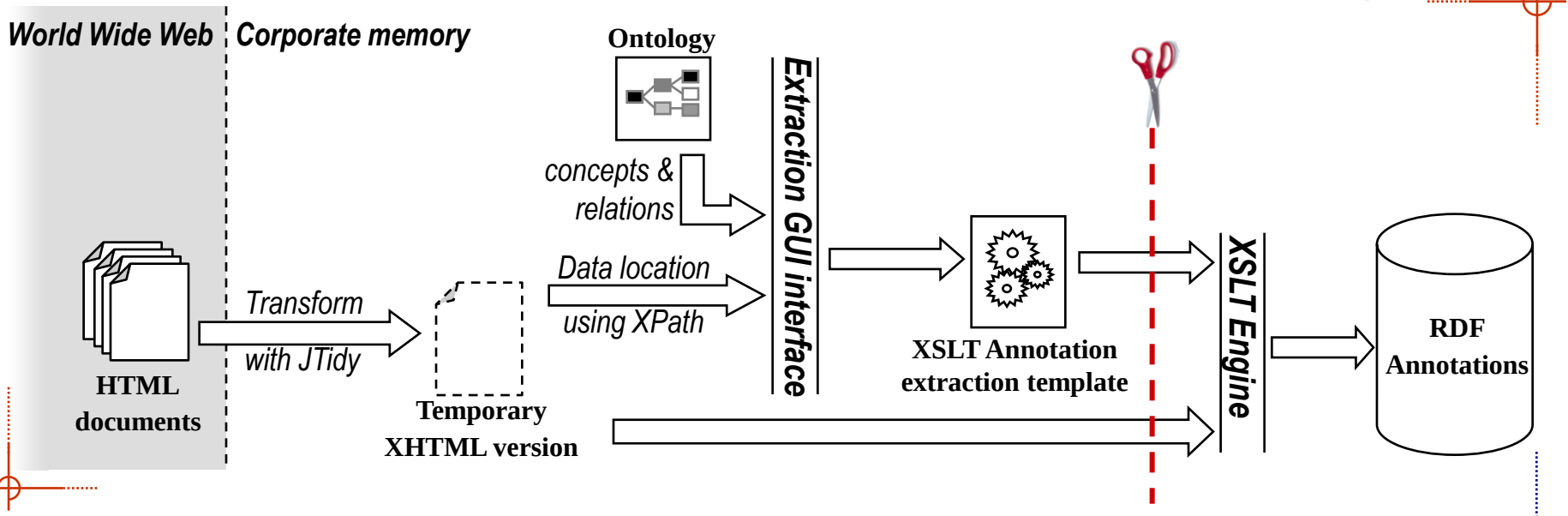
</InAnAsideToTheReviewers>

Two options for the annotation extraction

- On-the-fly conversion: generate annotations when needed up-to-date, no duplication, slow, connected
- Local archive maintenance: generate and maintain a base rapid, disconnected, needs monitoring, memory-consuming
- Second option: rapid & decouples intraweb from open-Web

XML-based approach:

principle



XML-based intraweb

- Ontology: XML syntax of RDF/S
- Web page: XHTML document
- Annotations: XML syntax of RDF

XSLT

- XML transformation language
- XML syntax: communication of extraction rules
- XPath expressions for data extraction

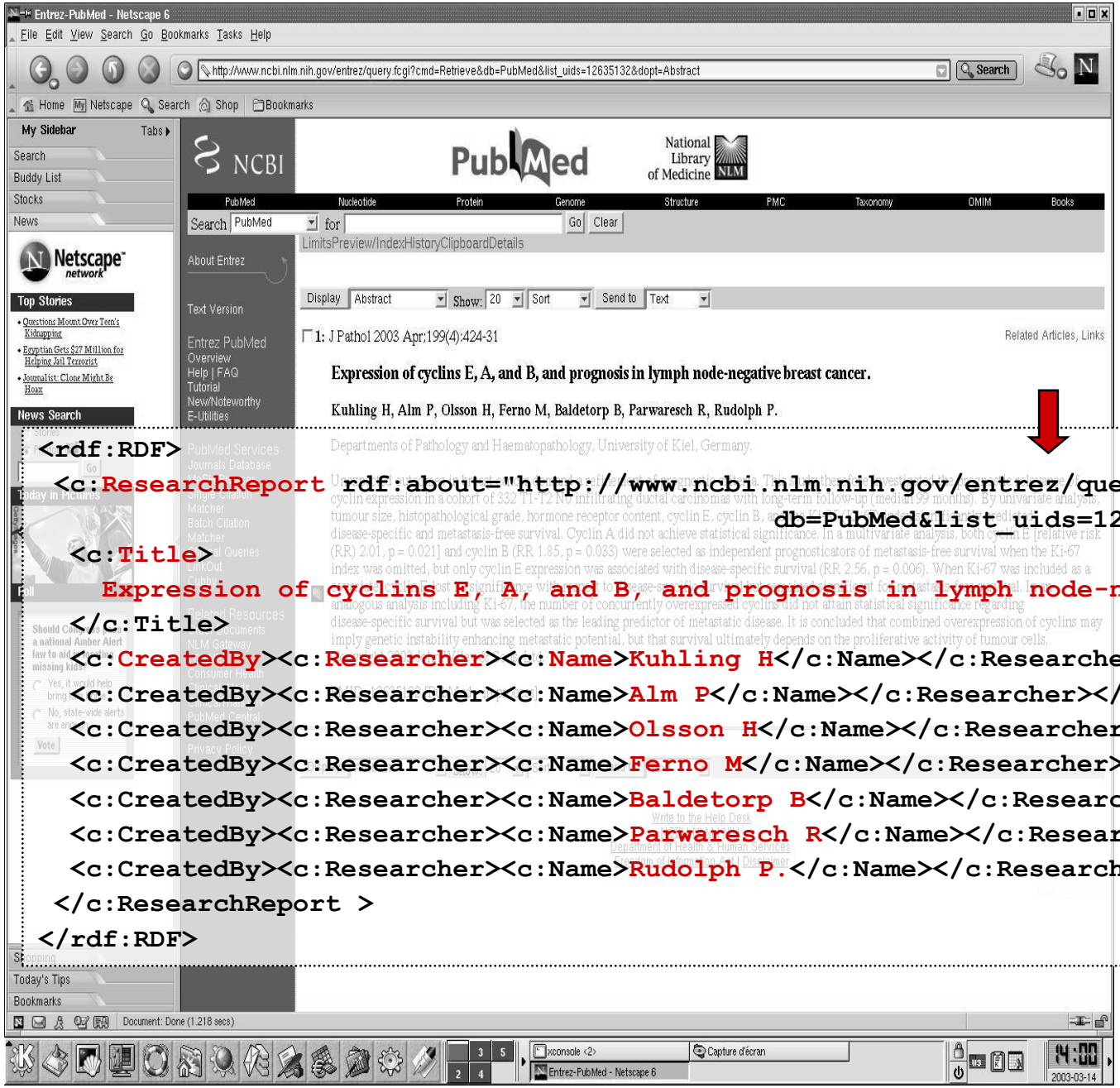
Template **construction**

- GUI **manipulation translation**
- **Built-in library** of extraction templates
  - e.g., list extraction, keyword mapping
  - Transparent, embedded, combined

example

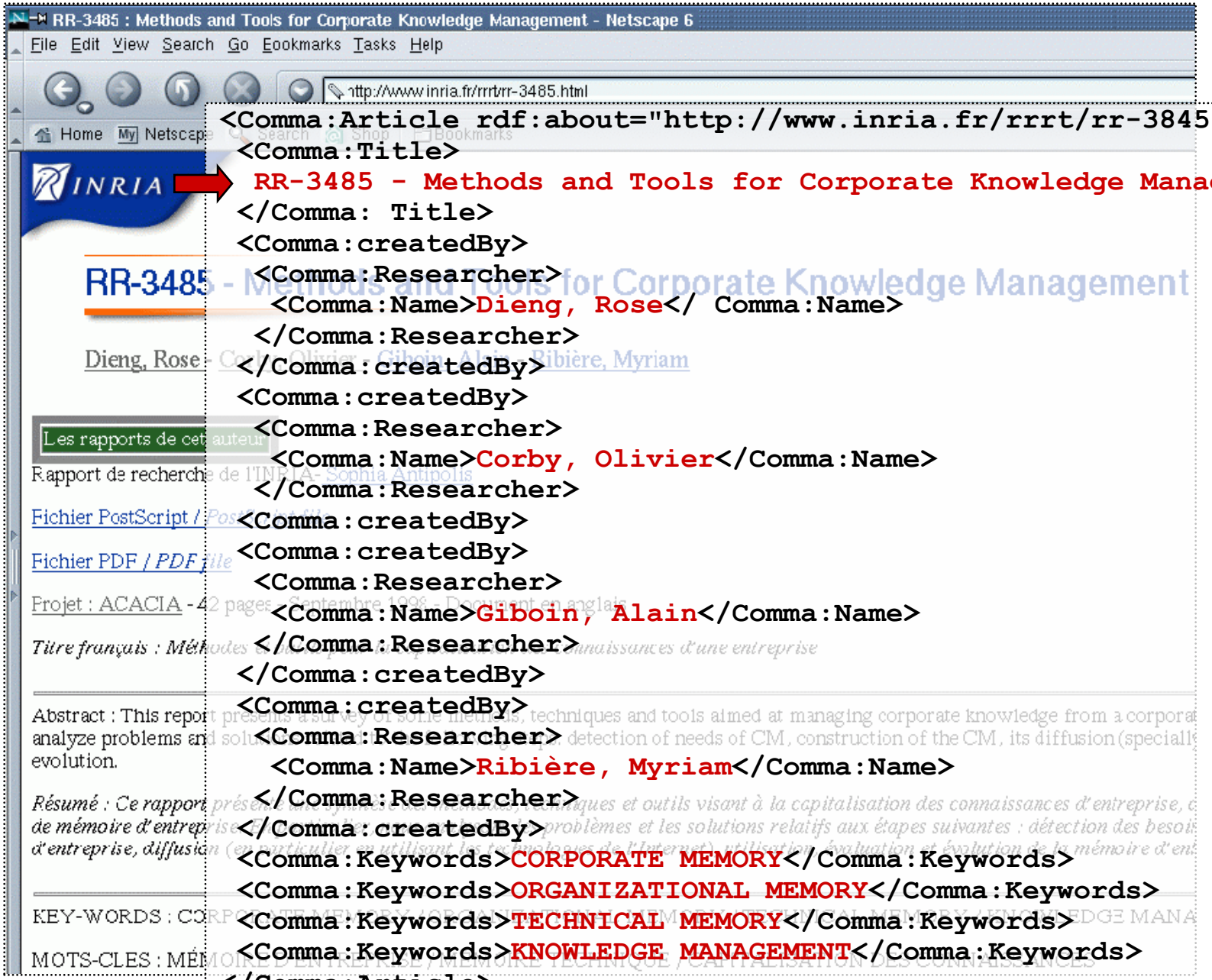
```
<xsl:template name="getListItem">
<xsl:param name="list" />
<xsl:param name="delimiter"/>
<xsl:param name="openning"/>
<xsl:param name="closing"/>
<xsl:choose>
 <xsl:when test ="$delimiter = 'br'">
  <xsl:for-each select="$list">
 <xsl:value-of select="$openning" disable-output-
                                          escaping="yes"/>
  <xsl:value-of select="normalize-space()" />
 <xsl:value-of select="$closing" disable-output-
                                          escaping="yes"/>
 </xsl:for-each>
 </xsl:when>
 <xsl:otherwise>
  <xsl:choose>
  <xsl:when test="string-length($list) = 0" />
  <xsl:otherwise>
    <xsl:choose>
       <xsl:when test="contains($list, $delimiter)">
      <xsl:value-of select="$openning"
                        disable-output-escaping="yes"/>
       <xsl:value-of select="concat(normalize-space
      (substring-before($list, $delimiter)),'&#10;')" />
      <xsl:value-of select="$closing"
                         disable-output-escaping="yes"/>
       </xsl:when>
       <xsl:otherwise>
     <xsl:value-of select="$openning"
                         disable-output-escaping="yes"/>
    <xsl:value-of select="concat(normalize-space
                                    ($list),'&#10;')" />
     <xsl:value-of select="$closing"
                         disable-output-escaping="yes"/>
       </xsl:otherwise>
    </xsl:choose>
   <xsl:call-template name="getListItem">
    <xsl:with-param name="list" select="substring-
                           after($list, $delimiter)" />
     <xsl:with-param name="delimiter"
                                   select="$delimiter" />
    <xsl:with-param name="openning"
                                    select="$openning" />
    <xsl:with-param name="closing" select="$closing" />
   </xsl:call-template>
  </xsl:otherwise>
 </xsl:choose>
 </xsl:otherwise>
</xsl:choose>
</xsl:template>
```

Example of extraction from the PubMed catalog



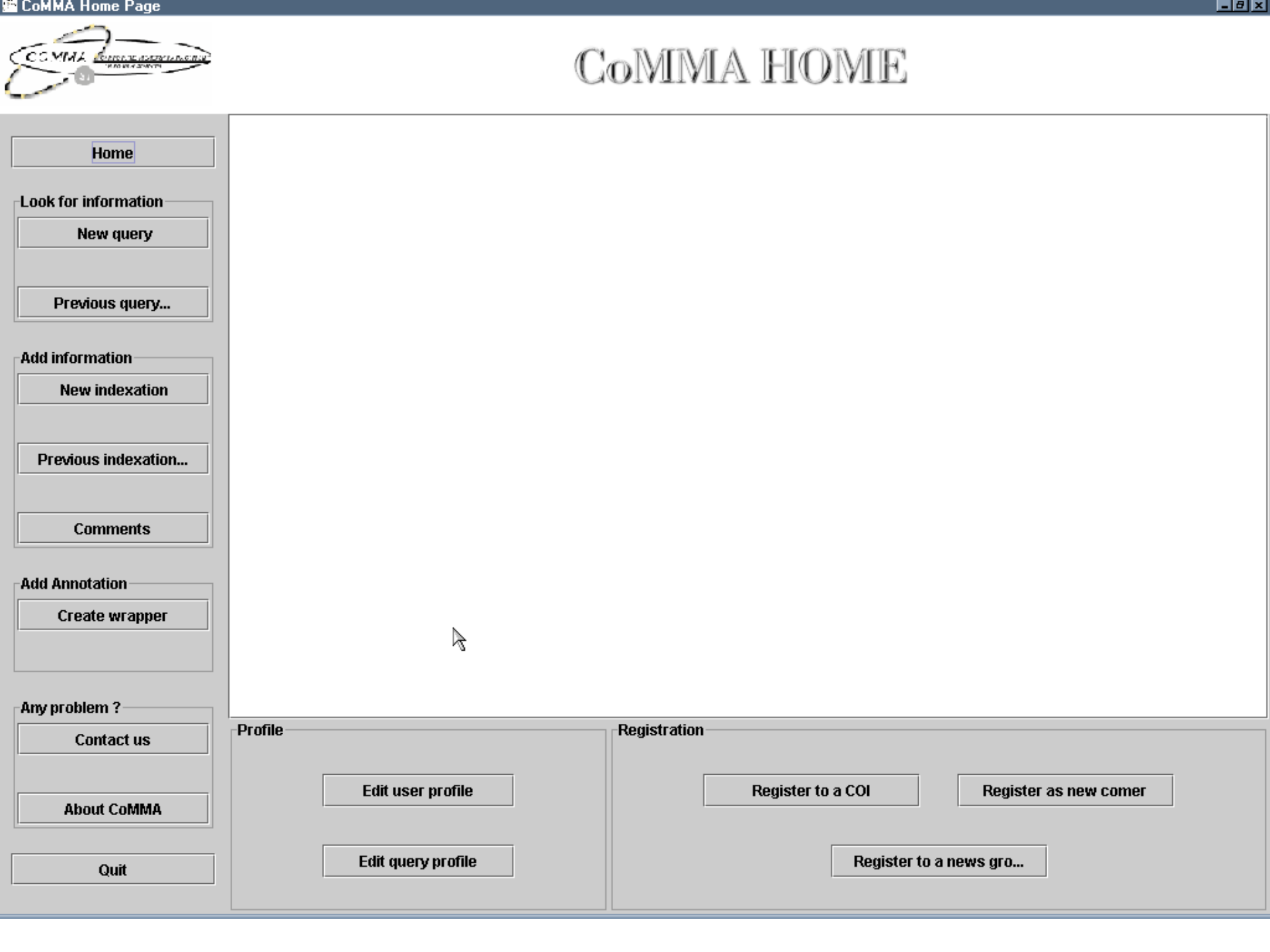

```
<rdf:RDF>
 <c:ResearchReport rdf:about="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&
                              db=PubMed&list_uids=12635132&dopt=Abstract\">
  <c:Title>
     Expression of cyclins E, A, and B, and prognosis in lymph node-negative breast cancer.
  </c:Title>
  <c:CreatedBy><c:Researcher><c:Name>Kuhling H</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Alm P</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Olsson H</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Ferno M</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Baldetorp B</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Parwaresch R</c:Name></c:Researcher></c:CreatedBy >
  <c:CreatedBy><c:Researcher><c:Name>Rudolph P.</c:Name></c:Researcher></c:CreatedBy >
 </c:ResearchReport >
</rdf:RDF>
```

Example of extraction from the INRIA librarian page




```
<Comma:Article rdf:about="http://www.inria.fr/rrrt/rr-3845.html">
 <Comma:Title>
  RR-3485 - Methods and Tools for Corporate Knowledge Management
 </Comma: Title>
 <Comma:createdBy>
  <Comma:Researcher>
   <Comma:Name>Dieng, Rose</ Comma:Name>
  </Comma:Researcher>
 </Comma:createdBy>
 <Comma:createdBy>
  <Comma:Researcher>
   <Comma:Name>Corby, Olivier</Comma:Name>
  </Comma:Researcher>
 <Comma:createdBy>
 <Comma:createdBy>
  <Comma:Researcher>
   <Comma:Name>Giboin, Alain</Comma:Name>
  </Comma:Researcher>
 </Comma:createdBy>
 <Comma:createdBy>
  <Comma:Researcher>
   <Comma:Name>Ribière, Myriam</Comma:Name>
  </Comma:Researcher>
 </Comma:createdBy>
 <Comma:Keywords>CORPORATE MEMORY</Comma:Keywords>
 <Comma:Keywords>ORGANIZATIONAL MEMORY</Comma:Keywords>
 <Comma:Keywords>TECHNICAL MEMORY</Comma:Keywords>
 <Comma:Keywords>KNOWLEDGE MANAGEMENT</Comma:Keywords>
</Comma:Article>
```

Modified the interface controller to add new GUI


CoMMA Home Page
CoMMA HOME
Home
Look for information
New query
Previous query...
Add information
New indexation
Previous indexation...
Comments
Add Annotation
Create wrapper
Any problem ?
Contact us
About CoMMA
Quit
Profile
Edit user profile
Edit query profile
Registration
Register to a COI
Register as new comer
Register to a news gro...

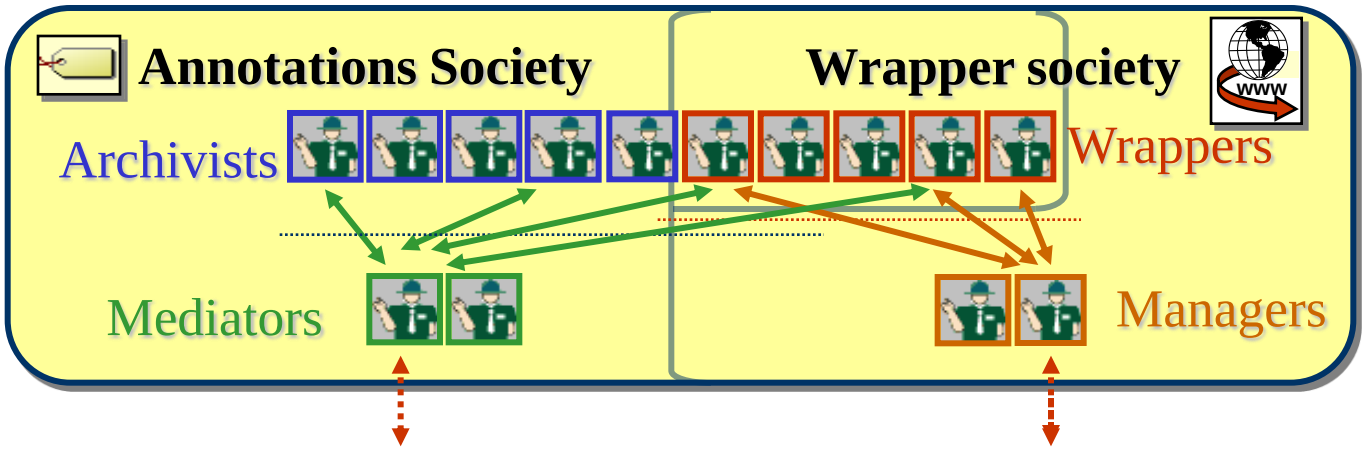
Modified the interface controller to add new GUI

The two roles of the **wrapper** society

- **Annotation wrapper archivists**: attached to 1 external source
- **Wrapper manager**: creates and manages wrappers
- **Hierarchical** society

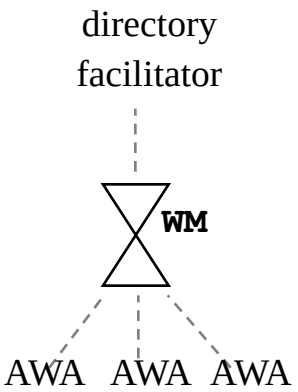
AWA: **2 (sub-)roles** involved in **2 societies**

- Annotation **Wrapper** role under Wrapper manager superv.
- Annotation **Archivist** role under Annotation Mediator superv.



Wrapper Manager

role model	Wrapper Manager role, part of the wrapper-dedicated society.
responsibilities	Wrappers management : create, recreate, monitor and destroy Annotation Wrapper Archivists
collaborators	Directory Facilitator, Interface Controller, Agent Management System, Annotation Wrapper Archivist
external interfaces	-
relationships	-
expertise	FIPA-Agent-Management Ontology, CoMMA-Ontology
interactions	Initiates FIPA-Request protocol to register and deregister itself with a Directory Facilitator. Responds to FIPA-Request protocol to create new AWA.
others	-

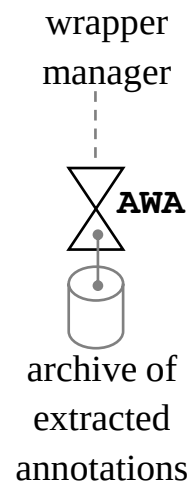


- **Contact point** for other sub-societies
- Currently only creates AWA

  but evolving toward managing wrapper's lifecycle

Annotation Wrapper Archivist

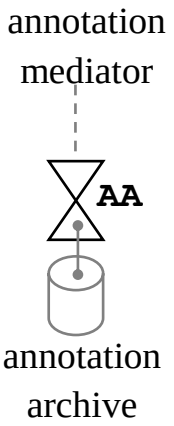
role model	Annotation Wrapper Archivist, part of the wrapper-dedicated society.
responsibilities	Wrap the Web source to generate RDF annotations. Store and manage these annotations. Search annotations to respond to queries.
collaborators	Directory facilitator, Wrapper Manager, Annotation Mediator
external interfaces	External Web site access and monitoring, (X)HTML manipulation, XSLT manipulation, RDF annotation manipulation interface
relationships	includes Annotation Archivist Role
expertise	annotation extraction query solving
interactions	Responds to FIPA-Request protocol to wrap a web site. Responds to FIPA-Query protocol to solve queries. Respond to FIPA-Contract-net to refuse archiving.
others	-



- **Extraction** and **maintenance** of a base of annotations
- Makes annotations available for **query solving**

Existing **Archivist** role (manage a local archive)

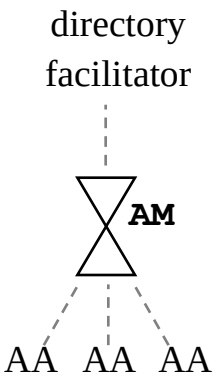
role model	Annotation Archivist role in the Annotation-dedicated society
responsibilities	store and query the annotations of the memory
collaborators	Directory facilitator, Annotation Mediator, Ontology Archivist
external interfaces	RDF annotation manipulation interface
relationships	also part of the roles in Corporate Model Archivist and User Profile Archivist
expertise	annotation archiving an querying
interactions	Query-Ref, Contract-Net; FIPA ACL
others	-



- Attached to & exploits **local base**
- **Answers** to query as much as it can with local knowledge
- Normally proposes **archiving** services
- One sub-behavior overwritten to refuse **archiving** services

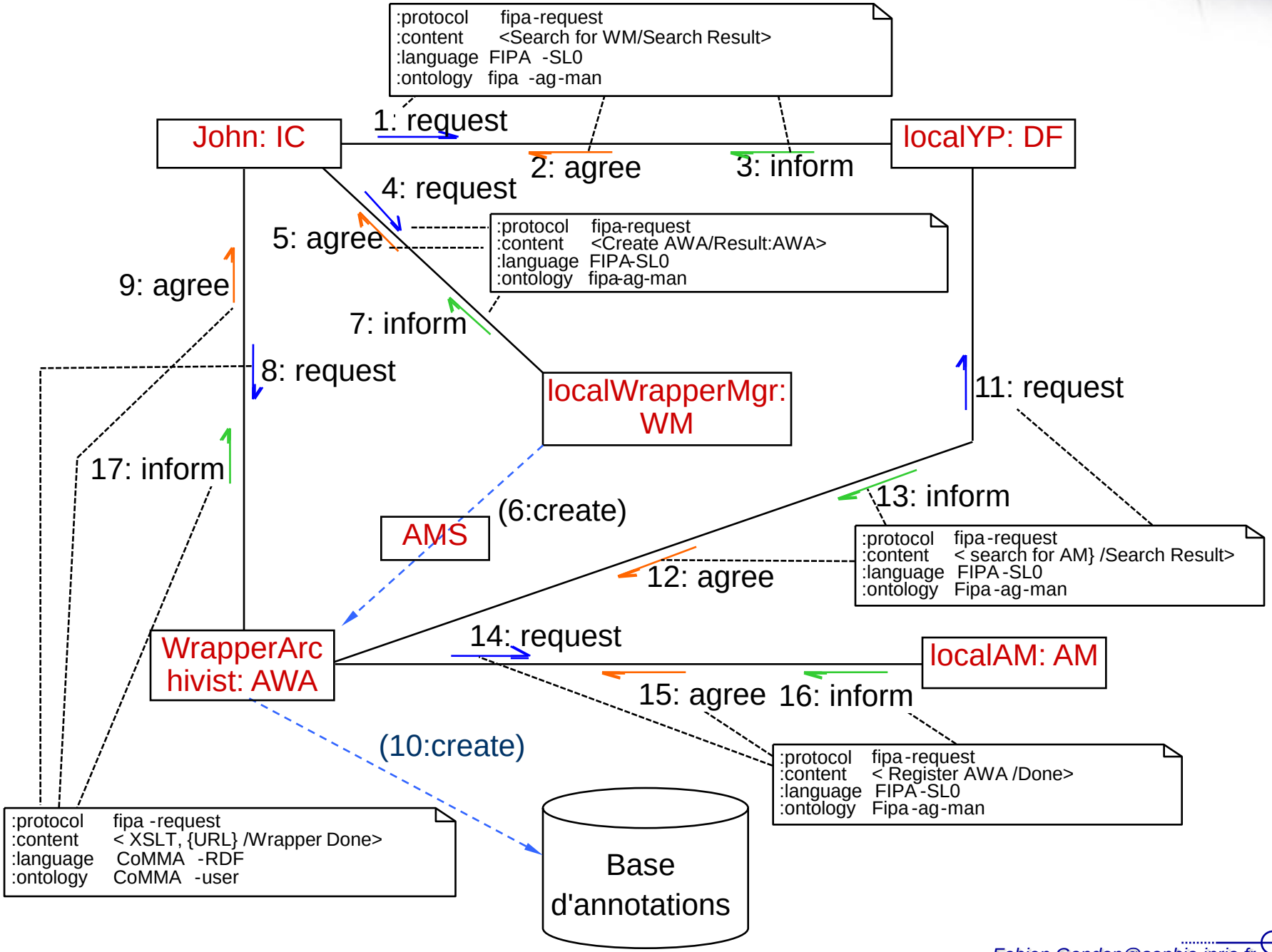
Existing **Annotation Mediator** **(untouched)**

role model	Annotation Mediator role in the Annotation-dedicated society
responsibilities	handle distribution of annotations over the archivists both for new annotation submissions and query solving processes
collaborators	Directory facilitator, User Profile Manager, Ontology Archivist, Annotation Archivist, Corporate Model Archivist
external interfaces	RDF annotation manipulation interface
relationships	-
expertise	query and submission management
interactions	Query-Ref, Contract-Net, Subscribe, Request; FIPA ACL
others	-



- **Contact point** for other sub-societies
- Supervising **distribution** of tasks for query solving
- **Allocating** a new annotation to an archive (semantic dist.)
- **Notifying** arrival new annotations to trigger push functions

Wrapping interaction protocol (acquaintance graph and interactions)



- Tested on **3 different libraries**: research reports of INRIA, technical reports of SCS – CMU, Library of medicines of MedLINE (correctly annotated)
- Generate and feed a large annotation base for performance evaluation of CORESE search engine
- **Improvements**:
  - **Relative XPaths, machine learning**, **multiple sources**
  - **Wrapper manager** role: manage the Wrappers Archivists life cycle (kill, restart, alerts to administrator, *etc.*)
  - **Annotation Wrapper Archivist** role: registration services to notify changes in the structure or content, automatic adjustment of XPath when the structure changes
- Applies to XML documents in general, including: RDF, XTopic, DAML+OIL, OWL, *etc.*
- General interest: **flexible** and **modular**, **customization** of sub-behaviors info. agents, exchange **procedural knowledge**, **standard**

- More about current & future work on **Wrapper**?
  mailto:TuanDung.Cao@sophia.inria.fr
- More about **CoMMA**?
  See PhD Dissertation “Distributed Artificial Intelligence and Knowledge Management: ontologies and multi-agent systems for a corporate semantic web”
  http://www-2.cs.cmu.edu/afs/.cs.cmu.edu/Web/People/fgandon/
  mailto:Fabien.Gandon@sophia.inria.fr / @cs.cmu.edu
- Current work at C.M.U. **“*my*Campus”**
  semantic web services & agents for context-aware nomadic PDA-based access to campus intraweb
  http://mycampus.sadehlab.cs.cmu.edu/project/

- Thanks