

FROM XML TO RDF IN THE ORLANDO PROJECT

John Simpson

Postdoctoral Fellow, University of Alberta
INKE / Text Mining & Visualization for Literary History
@symulation

Susan Brown

Professor, University of Alberta & University of Guelph
INKE / Text Mining & Visualization for Literary History
@susanirenebrown

With thanks to:

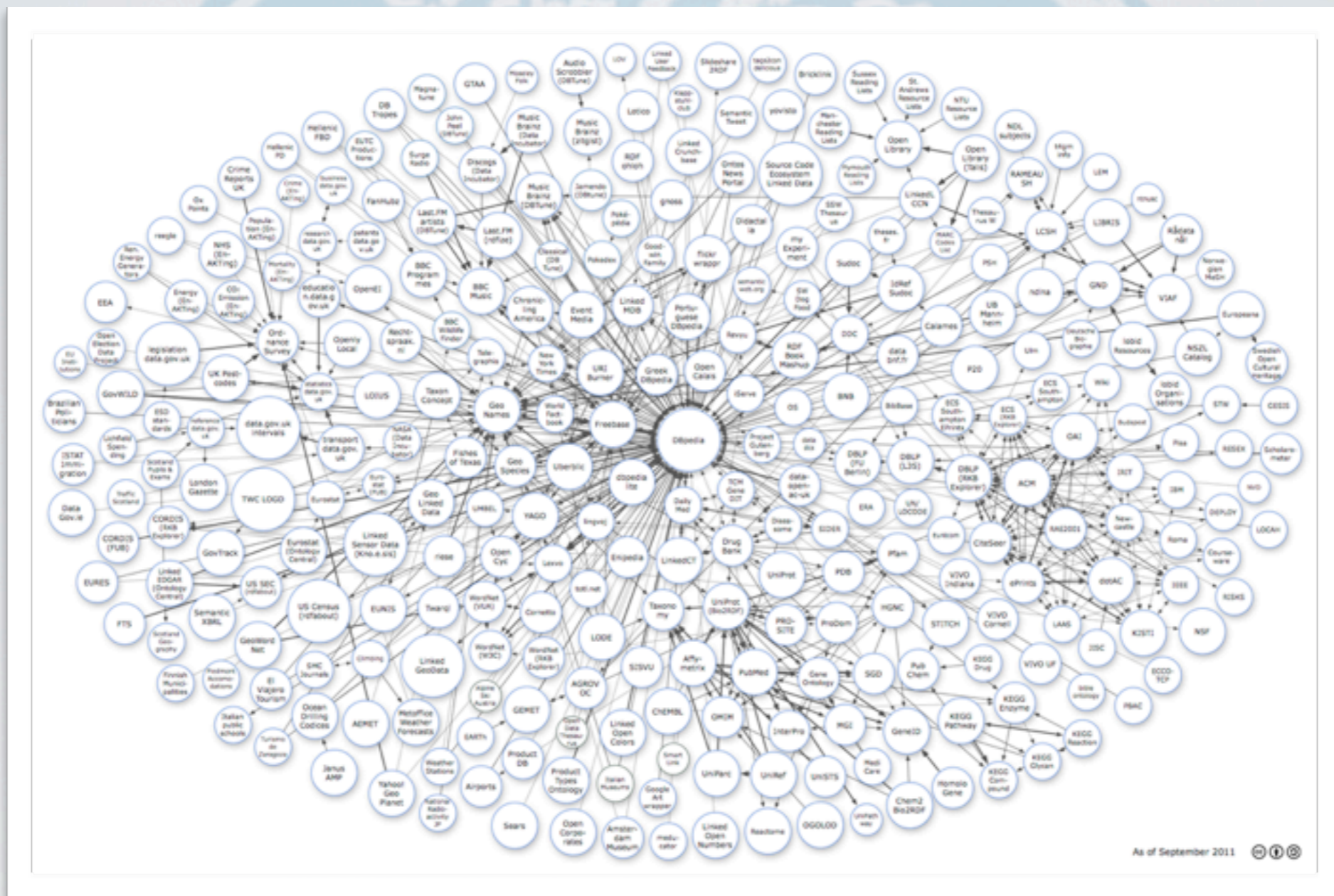
Jentery Sayers, Jon Adèle Barclay, and other members of the INKE Modeling & Prototyping team.

Members of Orlando, including editors Patricia Clements and Isobel Grundy.

Jeffery Antoniuk and other members of the Canadian Writing Research Collaboratory project (cwrc.ca)

- Introduce ourselves and our collaborators (Orlando and INKE)
- Paper offers a summary of technical challenges associated with the translation of content from rich XML in a boutique digital humanities project–The Orlando Project–into RDF.
- Road map:
 - quickly summarize Orlando and the motivation for LOD/semantic web/RDF (Susan) 3 minutes
 - what we did, how we did it, and why we did it that way (John) 4 minutes
 - challenges that have arisen (Susan) 3 minutes
 - other work we are doing and conclusions (John) 2 minutes
-

WHAT IS LINKED OPEN DATA?



Promise of:

Inference Based Search
Targeted Data Access
Graph Based Reasoning

- LOD is: Network of semantic connections across digital repositories, typically using RDF.
- Tim Berner's Lee calls it "The semantic web done right"
- See the "obligatory" LOD map (most recent version, 2011)
- LOD comes with some attractive promises:
 - Inference based search
 - Targeted data access
 - Graph based reasoning

ORLANDO: WOMEN'S WRITING IN THE BRITISH ISLES

Online cultural history generated from the lives and works of **over 1200 writers**, and for readers with an interest in literature, women's writing, or cultural history more generally. With over **eight million words of text**, it is full of interpretive information on women, writing, and culture.

Orlando currently features **1025 British women writers**--listed twice in cases of multiple, shifting, or contested nationality--; **13,607 free-standing chronology entries**; **26,278 bibliographical listings**; **2,499,869 tags**; **8,075,869 words (exclusive of tags)**

<http://www.arts.ualberta.ca/orlando/>

Share a little about its historical and current importance for the digital humanities

WHY THE TRANSITION?

For Orlando

- Extract new insights
- Easier linking with other resources
- As a case study in conversion for others

For DH

- Overcome Silos
- Public Face on Work
- Contribute to emerging semantic web

- This is really more explication of the promises at the bottom of SLIDE 2
- how does this advance Orlando?
 - hope to extract new insights that we can't get from xml-aware search, through ability to use sparql queries/inference engines and by visualizing the RDF graph may yield insights: e.g. who of the people associated with the poetess movement are key nodes? who are the people who are the most important links between the emergent feminist movement and the major periodicals of the mid-nineteenth century?
 - hope that it will become easier to link Orlando with other resources, e.g. to create an interface that combines digitized primary texts such as novels or poems with commentary from Orlando about those texts.
 - Provide a roadmap for others
- how does this help DH generally?
 - overcoming siloage; public-facing work; adding humanities insights to the emergent semantic web
 - tackling the sorts of challenges we outline here will improve the semantic web itself by making it more attentive to the gnarly problems of data representing history, lives, culture

TRANSITION CHALLENGES



Most “off-the-shelf” ontologies are not meant for humanities based data

Complicated relationships already embedded in both the XML and original prose

XML is nested, hierarchical, and undirected while RDF is flat and directed

- Three major challenges:
 - most standard ontologies are not meant for humanities data
 - embedded relationships are really complicated
 - XML and RDF each capture different sorts of properties best

WHAT DOES THE TRANSITION LOOK LIKE?

```
<ENTRY PERSON="BRWWRITER" SEX="FEMALE"
ID="balfcl"><BIOGRAPHY SEX="FEMALE" PERSON="BRWWRITER"
ID="balfcl-b.sgm"> <DATASTRUCT><DATAITEM><BIRTHNAME>
<GIVEN>Clara</GIVEN><SURNAME>Lucas</SURNAME></
BIRTHNAME></DATAITEM><DATAITEM><MARRIED
WROTEORPUBLISHEDAS="WROTEPUBLISHEDASYES"
REG="Balfour, Clara Lucas">Balfour</MARRIED><BIBCITS><BIBCIT
PLACEHOLDER="FC" DBREF="107913"/><BIBCIT
PLACEHOLDER="BLC" DBREF="2052"/></BIBCITS></
DATAITEM><DATAITEM><INDEXED>Clara Lucas Balfour</INDEXED></
DATAITEM></DATASTRUCT></DIV2></PERSONNAME></DIV1><DIV1
ID="b--balfcl--0--DIV1--2"><HEADING>Early Years</
HEADING><BIRTH><DIV2 ID="b--balfcl--0--DIV2--2"><CHRONSTRUCT
RELEVANCE="SELECTIVE"
CHRONCOLUMN="BRITISHWOMENWRITERS" RESP="SYS" ID="b--
balfcl--0--CHRONSTRUCT--1"><DATE VALUE="1808-12-21">21
December 1808</DATE><CHRONPROSE><NAME
STANDARD="Balfour, Clara">Clara Lucas</NAME> (later <NAME
STANDARD="Balfour, Clara">CB</NAME>) was born in the
<PLACE><REGION>New Forest</REGION><GEOG REG="England"/></
PLACE> in <PLACE><REGION>Hampshire</REGION><GEOG
REG="England"/></PLACE>.</CHRONPROSE><BIBCITS><BIBCIT
PLACEHOLDER="FC" DBREF="107913"/><BIBCIT
PLACEHOLDER="DNB" DBREF="1759"/></BIBCITS></
CHRONSTRUCT><SHORTPROSE><P ID="b--balfcl--0--
P--1"><BIRTHPOSITION POSITION="ONLY">She was an only child.</
BIRTHPOSITION><BIBCITS><BIBCIT PLACEHOLDER="FC"
DBREF="107913"/></BIBCITS></P></SHORTPROSE></DIV2></
BIRTH></DIV1><SHORTPROSE><P ID="b--balfcl--0--P--2">Her mother,
Sarah, <QUOTE DIRECT="Y">a woman of much intellectual power,</
QUOTE><BIBCITS><BIBCIT PLACEHOLDER="DNB" DBREF="1759"/
></BIBCITS> had apparently been tricked into a marriage that was
bigamous. The circumstances are obscure, but she disappeared from her
daughter's life when Clara was very young, to reappear after Clara's
father died.<BIBCITS>
```

- Snippet of the XML for an entry on Clara Balfour.
 - Point out her birthday, gender, name, something about her mother
- Show format after change.

WHAT DOES THE TRANSITION LOOK LIKE?

```
<rdf:Description rdf:about="Balfour,  
Clara"><ex:hasSex rdf:resource="FEMALE"/  
></rdf:Description>
```

```
<rdf:Description rdf:about="Balfour,  
Clara"><ex:givenName rdf:resource="Clara"/  
></rdf:Description>
```

```
<rdf:Description rdf:about="Balfour,  
Clara"><ex:surname rdf:resource="Lucas"/></  
rdf:Description>
```

```
<rdf:Description rdf:about="Balfour,  
Clara"><ex:dateOfBirth  
rdf:resource="1808-12-21"/></  
rdf:Description>
```

- Snippet of the XML for an entry on Clara Balfour.
 - Point out her birthday, gender, name, something about her mother
- Show format after change.

RDF SYNTAX

```
<rdf:Description rdf:about="Balfour, Clara">
```

```
<ex:wasBornOn rdf:resource="1808-12-21"/>
```

```
</rdf:Description>
```

subject

predicate

object

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Show subject, predicate, object nature
- A

RDF SYNTAX

```
<rdf:Description rdf:about="Balfour, Clara">
```

```
<ex:wasBornOn rdf:resource="1808-12-21"/>
```

```
</rdf:Description>
```

Balfour, Clara

subject

predicate

object

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Show subject, predicate, object nature

RDF SYNTAX

```
<rdf:Description rdf:about="Balfour, Clara">
```

```
<ex:wasBornOn rdf:resource="1808-12-21"/>
```

```
</rdf:Description>
```

Balfour, Clara

wasBornOn

subject

predicate

object

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Show subject, predicate, object nature

RDF SYNTAX

```
<rdf:Description rdf:about="Balfour, Clara">
```

```
<ex:wasBornOn rdf:resource="1808-12-21"/>
```

```
</rdf:Description>
```

Balfour, Clara

wasBornOn

1808-12-21

subject

predicate

object

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Show subject, predicate, object nature

MAIN TRANSLATION OPTIONS



- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

•

MAIN TRANSLATION OPTIONS

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited
use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

•

MAIN TRANSLATION OPTIONS

Python + REGEX

Fast Prototyping

Light Replication

RegEx useful
elsewhere

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited
use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

MAIN TRANSLATION OPTIONS

Python + REGEX

Fast Prototyping

Light Replication

RegEx useful
elsewhere

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited
use elsewhere

XQUERY

Fast Prototyping

Light Replication

XQUERY has limited
use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

MAIN TRANSLATION OPTIONS

Python + REGEX

Fast Prototyping

Light Replication

RegEx useful
elsewhere

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited
use elsewhere

XQUERY

Fast Prototyping

Light Replication

XQUERY has limited
use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

MAIN TRANSLATION OPTIONS

Python + REGEX

Fast Prototyping

Light Replication

RegEx useful
elsewhere

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited
use elsewhere

XQUERY

Fast Prototyping

Light Replication

XQUERY has limited
use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

MAIN TRANSLATION OPTIONS

Python + REGEX

Fast Prototyping

Light Replication

RegEx useful elsewhere

XSLT

Fast Prototyping

Heavy Replication

XSLT has limited use elsewhere

XQUERY

Fast Prototyping

Light Replication

XQUERY has limited use elsewhere

- Python & Regex, XSLT, or XQUERY
- Could have used any, but already knew python and regular expressions and it had the most flexibility
 - option to use XPATH in Python remained open
- Ended up with a script that reads nested regex statements

RESULTS

2,499,896 XML TAGS
+ 100 LINES OF PYTHON
+ 80 LINES OF REGEX

542,710 RDF TRIPLES

Relatively little work over creating the XML for a fairly significant output

LIMITATIONS

FALSE POSITIVES

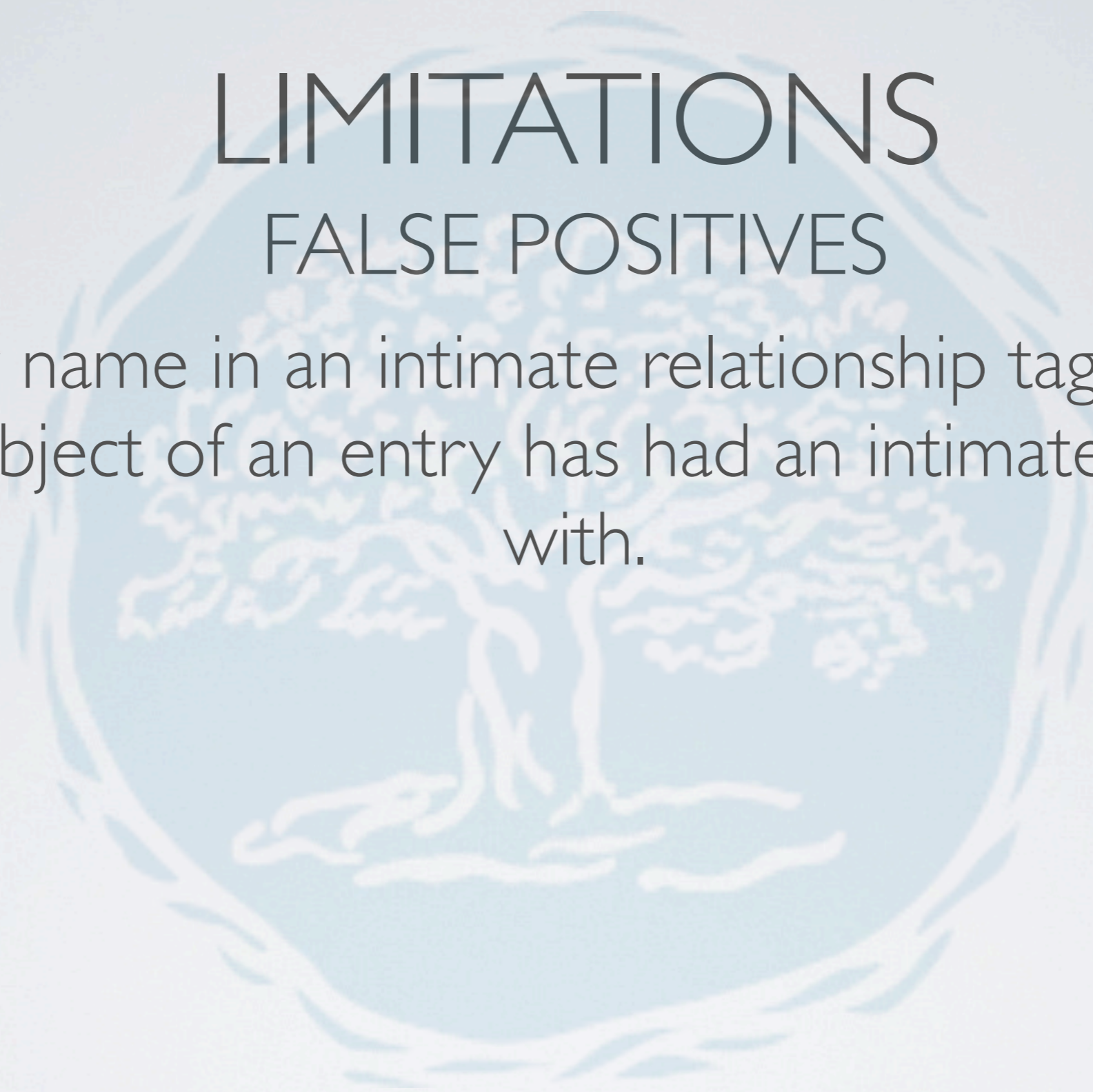
Not every name in an intimate relationship tag is someone that the subject of an entry has had an intimate relationship with.

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Expect that every person in an RELATIONSHIP tag would have at least a direct relationship to the subject of the entry.
- Often true, but we can't assume it to be so.

LIMITATIONS

FALSE POSITIVES

Not every name in an intimate relationship tag is someone that the subject of an entry has had an intimate relationship with.



LIMITATIONS

FALSE POSITIVES

Not every name in an intimate relationship tag is someone that the subject of an entry has had an intimate relationship with.

<INTIMATERELATIONSHIP>However, the intense jealousy that had early affected the friendship persisted. Around 1846 <NAMESTANDARD=Jewsbury, Geraldine > GJ </NAME> began her friendship with the American actress <NAME STANDARD=Cushman, Charlotte > Charlotte Cushman </NAME> , who was then visiting Manchester. Much to the frustration of Carlyle, Jewsbury began to write affectionate letters to Cushman, with whom she also reportedly <QUOTE DIRECT=Y > swore eternal friendship. 70 Years later <NAME STANDARD=Jewsbury, Geraldine > GJ </NAME> was enraptured by another actress, <NAME STANDARD=Faucit, Helen > Helena Faucit Martin </NAME> , whom she had seen perform many of <NAME STANDARD=Shakespeare, William > Shakespeare </NAME> 's heroines. She sent Martin adoring letters.</INTIMATERELATIONSHIP>

Case of Geraldine Jewsbury

LIMITATIONS

FALSE POSITIVES

Not every name in an intimate relationship tag is someone that the subject of an entry has had an intimate relationship with.

<INTIMATERELATIONSHIP>However, the intense jealousy that had early affected the friendship persisted. Around 1846 <NAMESTANDARD=Jewsbury, Geraldine > GJ </NAME> began her friendship with the American actress <NAME STANDARD=Cushman, Charlotte > Charlotte Cushman </NAME> , who was then visiting Manchester. Much to the frustration of Carlyle, Jewsbury began to write affectionate letters to Cushman, with whom she also reportedly <QUOTE DIRECT=Y > swore eternal friendship. 70 Years later <NAME STANDARD=Jewsbury, Geraldine > GJ </NAME> was enraptured by another actress, <NAME STANDARD=Faucit, Helen > Helena Faucit Martin </NAME> , whom she had seen perform many of <NAME STANDARD=Shakespeare, William > Shakespeare </NAME> 's heroines. She sent Martin adoring letters.</INTIMATERELATIONSHIP>

Geraldine Jewsbury has not had an intimate relationship with Shakespeare.

Case of Geraldine Jewsbury

LIMITATIONS

MISSED CONNECTIONS

Relationships between co-located entities cannot be extracted.

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Would really like to be able to relate people within an entry but who the entry isn’t about.
- The XML mark-up as used doesn’t allow this

LIMITATIONS

MISSED CONNECTIONS

Relationships between co-located entities cannot be extracted.



LIMITATIONS

MISSED CONNECTIONS

Relationships between co-located entities cannot be extracted.

<NAME STANDARD=Cobbe, Frances Power > FPC </NAME> 's connections from home gave her introductions into the circles of US and British women living in Italy, including <NAME STANDARD=Hosmer, Harriet > Harriet Hosmer </NAME> (who became a close friend). She met <NAME STANDARD=Browning, Elizabeth Barrett > Elizabeth Barrett</NAME> and <NAME STANDARD=Browning, Robert > Robert Browning </NAME> , and <NAME STANDARD=Trollope, Theodosia > Theodosia </NAME> and <NAME STANDARD=Trollope, Thomas Adolphus > Thomas Adolphus Trollope </NAME> , <LIVESWITH > and may have lived for a time with <NAME STANDARD=Blagden, Isa > Isa Blagden</NAME></LIVESWITH>. She also met other writers with common interests, including <NAME STANDARD=Mackay, Robert William > Robert William Mackay </NAME>, with whom she travelled from Rome to Naples on her way to the east.

Case of Frances Power Cobbe

LIMITATIONS

MISSED CONNECTIONS

Relationships between co-located entities cannot be extracted.

<NAME STANDARD=Cobbe, Frances Power > FPC </NAME> 's connections from home gave her introductions into the circles of US and British women living in Italy, including <NAME STANDARD=Hosmer, Harriet > Harriet Hosmer </NAME> (who became a close friend). She met <NAME STANDARD=Browning, Elizabeth Barrett > Elizabeth Barrett</NAME> and <NAME STANDARD=Browning, Robert > Robert Browning </NAME> , and <NAME STANDARD=Trollope, Theodosia > Theodosia </NAME> and <NAMESTANDARD=Trollope, Thomas Adolphus > Thomas Adolphus Trollope </NAME> , <LIVESWITH > and may have lived for a time with <NAME STANDARD=Blagden, Isa > Isa Blagden</NAME></LIVESWITH>. She also met other writers with common interests, including <NAME STANDARD=Mackay, Robert William > Robert William Mackay </NAME>, with whom she travelled from Rome to Naples on her way to the east.

There is nothing in the XML mark-up that allows for the members of the circle to be directly related.

Case of Frances Power Cobbe

PROVENANCE



- given the limitations; need to make fact of automated extraction clear; could be done via meta-level tags in the file that holds the RDF, but this would be hidden from queries and inference agents. To avoid this, each triple is reified by assigning it a URI and then connecting its pieces with the predicates `rdf:object`, `rdf:predicate`, and `rdf:subject`.

PROVENANCE

Balfour, Clara wasBornOn → 1808-12-21



- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

PROVENANCE

Balfour, Clara $\xrightarrow{\text{wasBornOn}}$ 1808-12-21

Reified Triple

- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

PROVENANCE

Balfour, Clara $\xrightarrow{\text{wasBornOn}}$ 1808-12-21



- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

PROVENANCE

Balfour, Clara $\xrightarrow{\text{wasBornOn}}$ 1808-12-21



- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

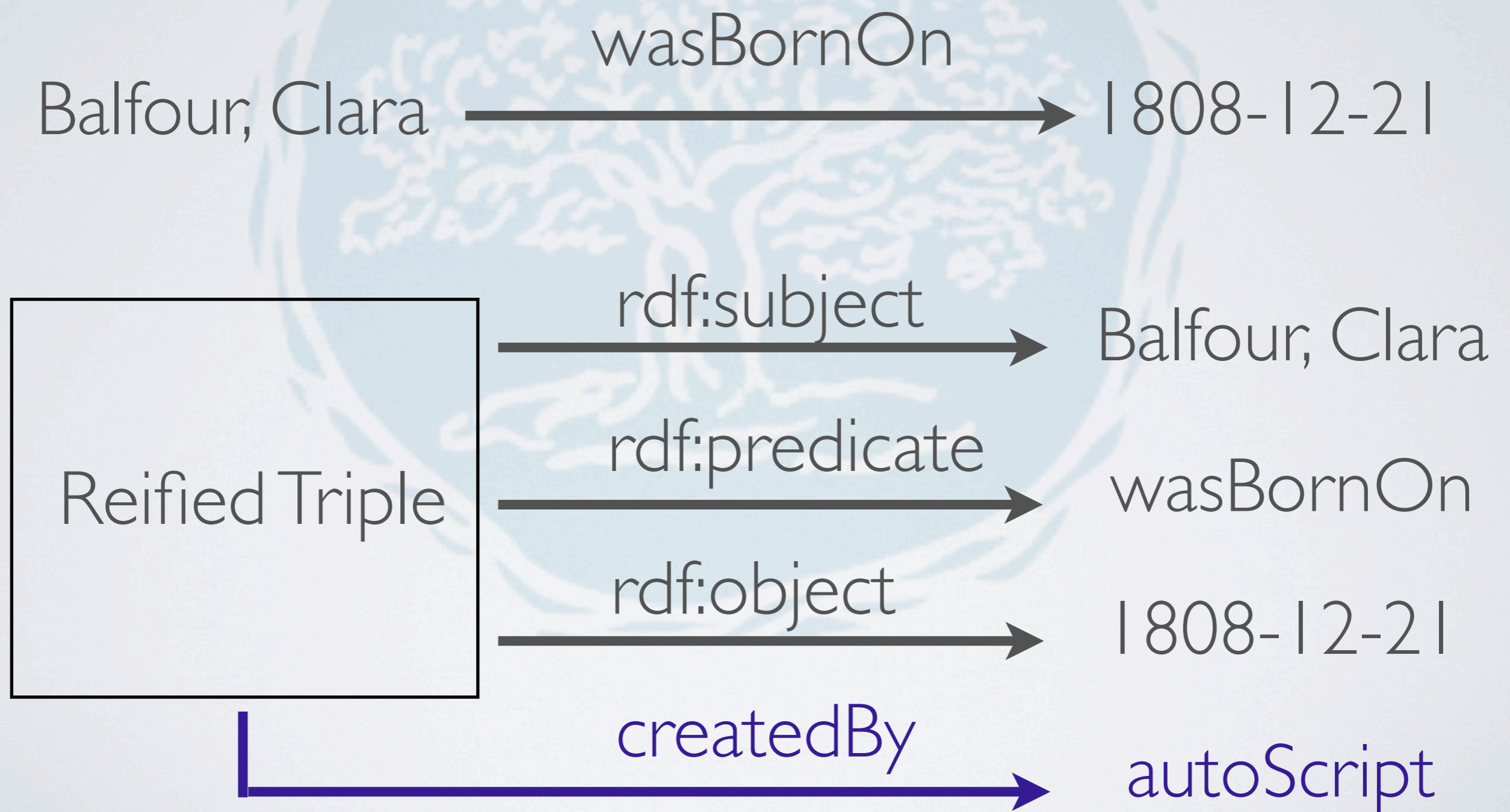
PROVENANCE

Balfour, Clara $\xrightarrow{\text{wasBornOn}}$ 1808-12-21



- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

PROVENANCE



- Really just one slide in concept but the use of “magic move” to make the text slide requires multiple slides
- Given the sorts of limitations that exist the tool will make errors.
- So we need to call out that any triples made this way came from an automatic script.
- To do this we reify the triple, making it a subject in its own right and making the original components of the triple objects.
- Can now predicate “created by” and have “auto script” as the object
- Of course, this makes it harder to query the triple-store

PARALLEL WORK

XML Conversion Scripts

- Two NER tools are being built by Denilson Barbosa and students at the U of A: SONEX and EXEMPLAR.
- These will process raw text (XML stripped) and provide RDF
- A way to get past some of the XML limitations
- XML to RDF tool will be used to help train it

PARALLEL WORK



XML
Conversion Scripts

Named Entity Recognition
SONEX
EXEMPLAR

- Two NER tools are being built by Denilson Barbosa and students at the U of A: SONEX and EXEMPLAR.
- These will process raw text (XML stripped) and provide RDF
- A way to get past some of the XML limitations
- XML to RDF tool will be used to help train it

PARALLEL WORK



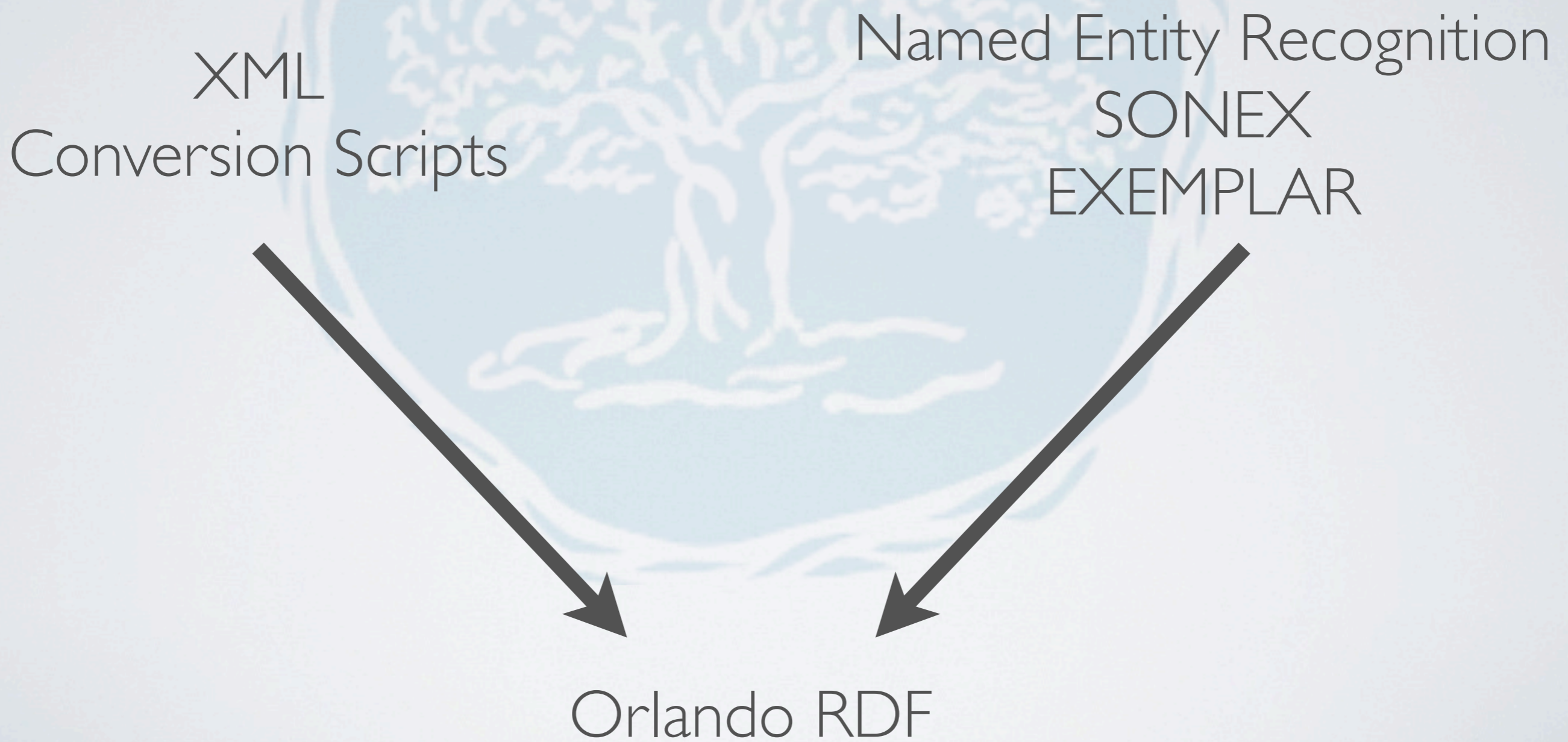
XML
Conversion Scripts

Named Entity Recognition
SONEX
EXEMPLAR

Orlando RDF

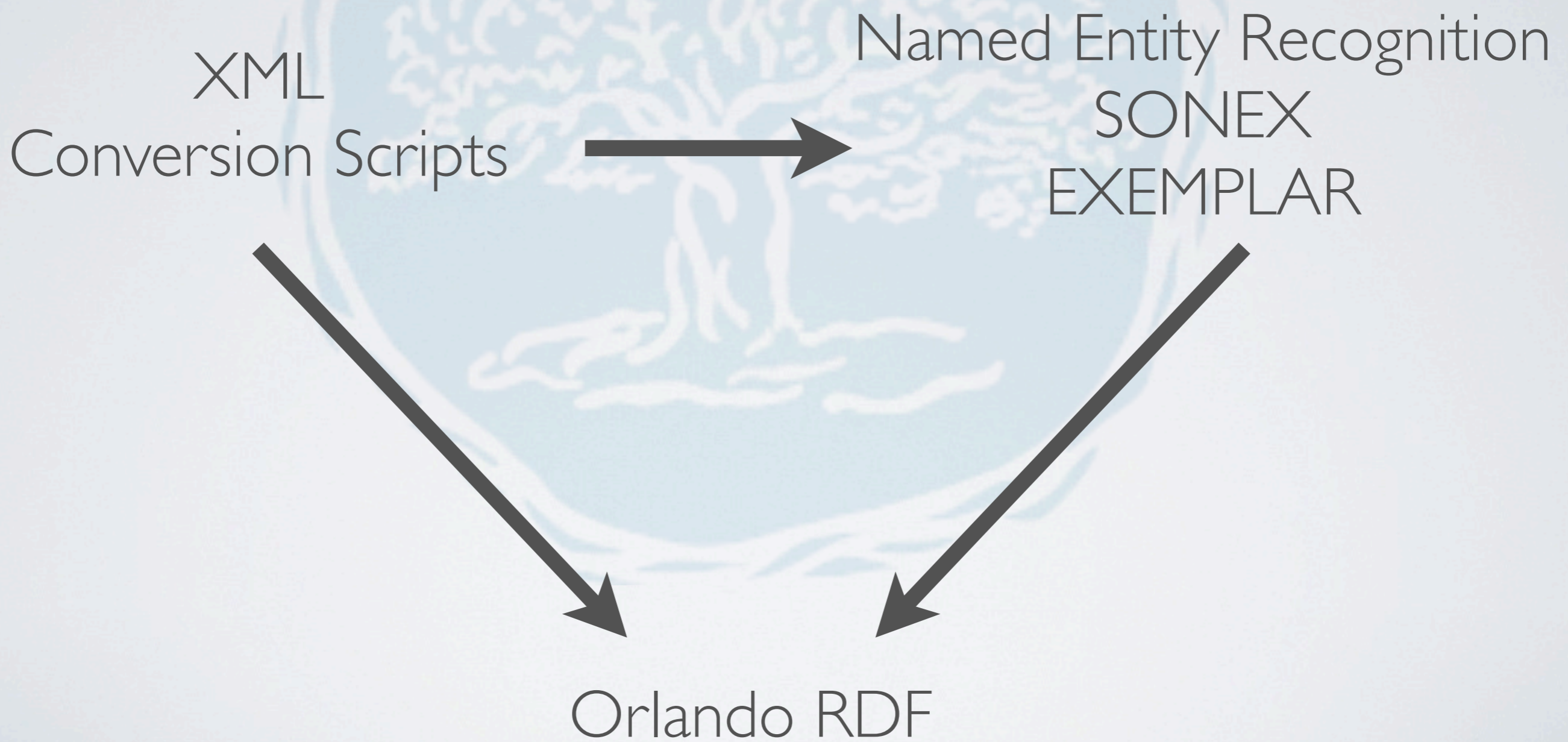
- Two NER tools are being built by Denilson Barbosa and students at the U of A: SONEX and EXEMPLAR.
- These will process raw text (XML stripped) and provide RDF
- A way to get past some of the XML limitations
- XML to RDF tool will be used to help train it

PARALLEL WORK



- Two NER tools are being built by Denilson Barbosa and students at the U of A: SONEX and EXEMPLAR.
- These will process raw text (XML stripped) and provide RDF
- A way to get past some of the XML limitations
- XML to RDF tool will be used to help train it

PARALLEL WORK



- Two NER tools are being built by Denilson Barbosa and students at the U of A: SONEX and EXEMPLAR.
- These will process raw text (XML stripped) and provide RDF
- A way to get past some of the XML limitations
- XML to RDF tool will be used to help train it

CONCLUSIONS



- Exploring this translation process pushes at the limits of formal and informal languages.
- This brings out the benefits and detriments of each and shows the nuances of our natural languages.
- As humanists we need to embrace this process even though it is messy as it reveals our own content in new ways and it keeps our voice in the broader community.

•

CONCLUSIONS

Conversion of XML to RDF is not a straightforward process.

- Exploring this translation process pushes at the limits of formal and informal languages.
- This brings out the benefits and detriments of each and shows the nuances of our natural languages.
- As humanists we need to embrace this process even though it is messy as it reveals our own content in new ways and it keeps our voice in the broader community.

•

CONCLUSIONS

Conversion of XML to RDF is not a straightforward process.

It is complicated by the inability of formal mark-up RDF and XML to capture all the relevant nuances and capacities of our natural languages.

- Exploring this translation process pushes at the limits of formal and informal languages.
- This brings out the benefits and detriments of each and shows the nuances of our natural languages.
- As humanists we need to embrace this process even though it is messy as it reveals our own content in new ways and it keeps our voice in the broader community.

•

CONCLUSIONS


Conversion of XML to RDF is not a straightforward process.

It is complicated by the inability of formal mark-up RDF and XML to capture all the relevant nuances and capacities of our natural languages.

The way forward for the humanities is to embrace such conversions so that our needs are part of the development process.

- Exploring this translation process pushes at the limits of formal and informal languages.
- This brings out the benefits and detriments of each and shows the nuances of our natural languages.
- As humanists we need to embrace this process even though it is messy as it reveals our own content in new ways and it keeps our voice in the broader community.

•



FROM XML TO RDF IN THE ORLANDO PROJECT

John Simpson

Postdoctoral Fellow, University of Alberta
INKE / Text Mining & Visualization for Literary History
@symulation

Susan Brown

Professor, University of Alberta & University of Guelph
INKE / Text Mining & Visualization for Literary History
@susanirenebrown

With thanks to:

Jentery Sayers, Jon Adèle Barclay, and other members of the INKE Modeling & Prototyping team.

Members of Orlando, including editors Patricia Clements and Isobel Grundy.

Jeffery Antoniuk and other members of the Canadian Writing Research Collaboratory project (cwrc.ca)

Just thanks and questions



This is just a placeholder in case we overshoot on the last slide