From monolithic XML for print/web to lean XML for data: realising linked data for dictionaries

Matt Kohl & Sandro Cirulli
Language Technologists
Oxford University Press (OUP)

7 June 2014

Introduction
Oxford University Press

OXFORD
UNIVERSITY PRESS

- **World-renowned dictionary publisher**
- **Licensing partner for lexical data**

Introduction
Shifts in Publishing

OXFORD
UNIVERSITY PRESS

- New trends & demands
- Emerging technologies & markets
- Importance of well-structured, semantically-rich data
- Speed!

Data Modelling
Our Current Dictionary Data Models

OXFORD UNIVERSITY PRESS

- **Print-oriented**: designed to capture dictionary layout
- **Monolithic**: one enormous document
- **Permissive**: continually loosened to accommodate new texts

Can't give us the flexibility we need

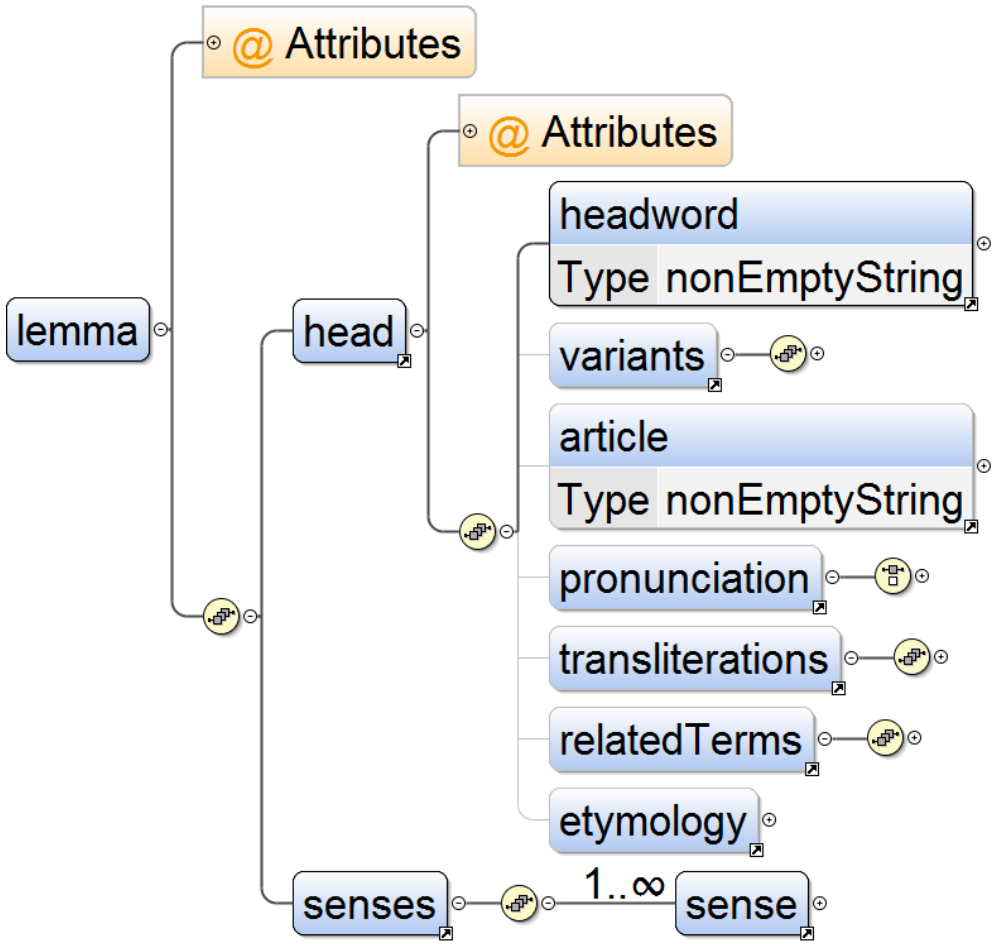
Data Modelling
Requirements

OXFORD UNIVERSITY PRESS

A new approach should:

- **Represent language concepts**, not layouts
- **Enable data reusability** for different products & services
- **Allow only one, clear way to model any given lexical item**

Data Modelling
The New Lexical Schema



Data Conversion
Moving Data into the Lexical Schema

OXFORD UNIVERSITY PRESS

Conversion Framework Requirements

- **Scalability**: convert 40+ data-sets
- **Standardization**: harmonize variation inside the data-sets
- **Modularity**: enable customization, slotting in & out of QA, etc.

Data Conversion
Tools

OXFORD
UNIVERSITY PRESS

- **XProc**
- **XSpec**
- **Schematron & XML Schema**
- **Jenkins CI**
- **Agile methodology**

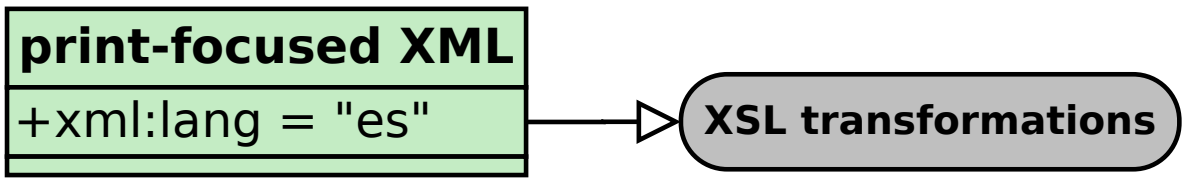
Data Conversion

Simplified XProc pipeline

OXFORD UNIVERSITY PRESS

print-focused XML
+xml:lang = "es"

Data Conversion
Simplified XProc pipeline

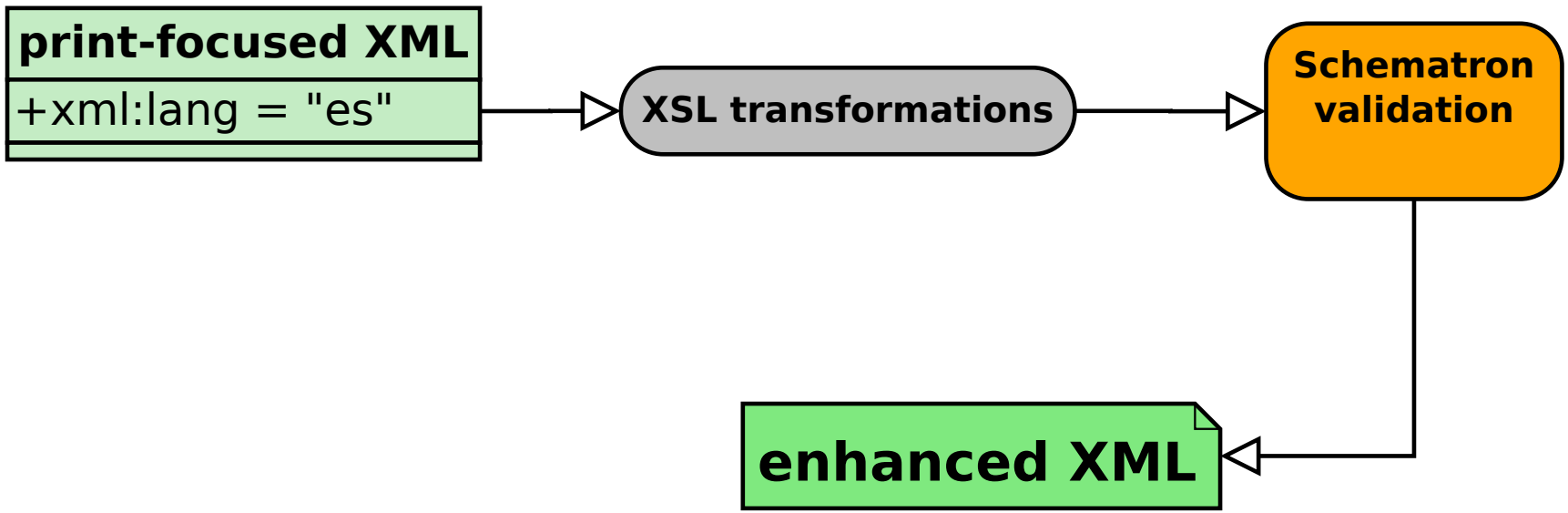


Data Conversion

Simplified XProc pipeline

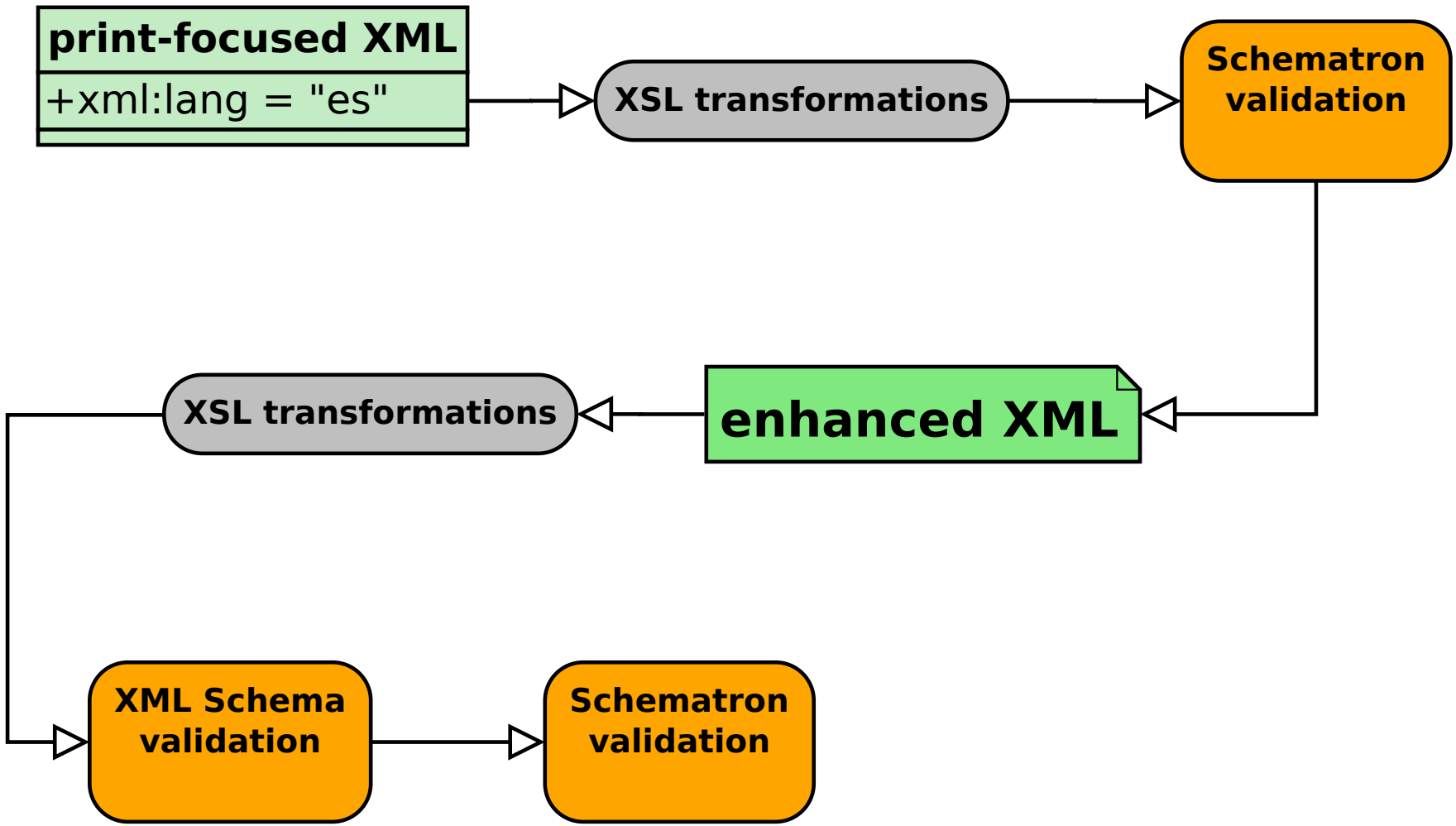
Data Conversion

Simplified XProc pipeline



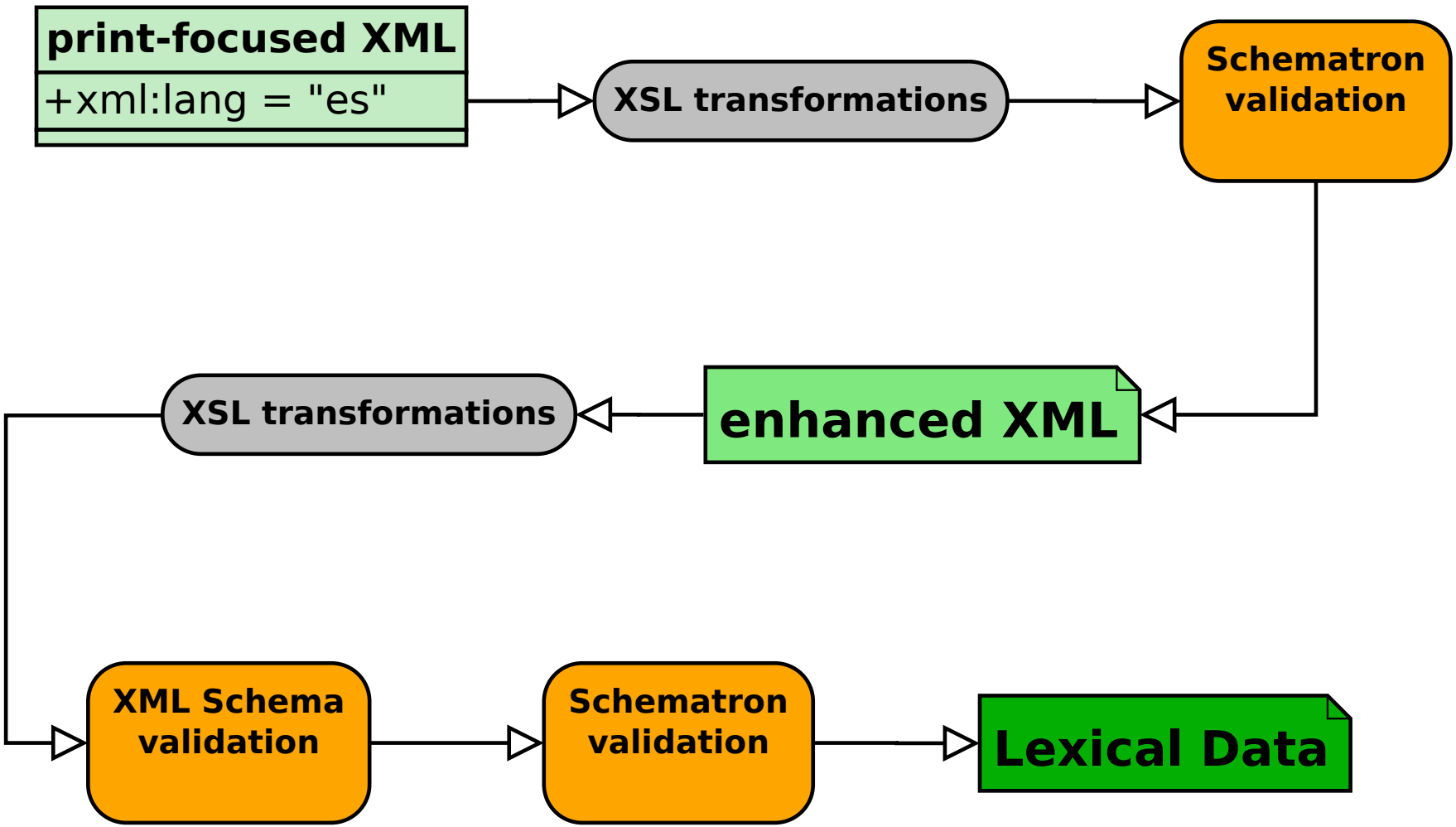
Data Conversion

Simplified XProc pipeline



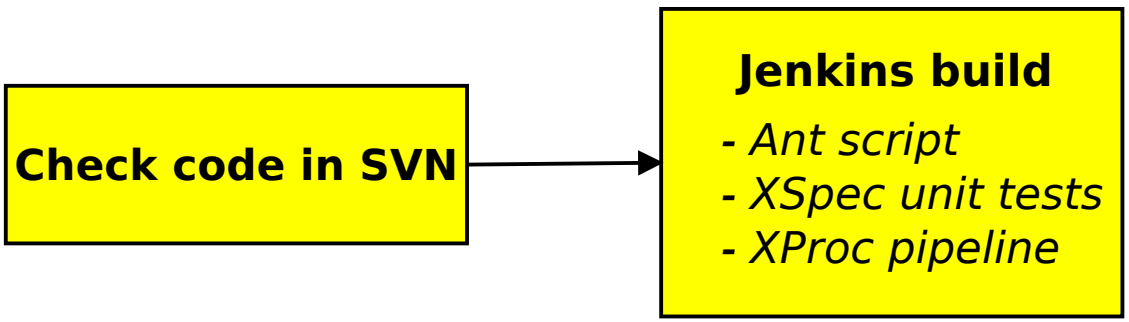
Data Conversion

Simplified XProc pipeline



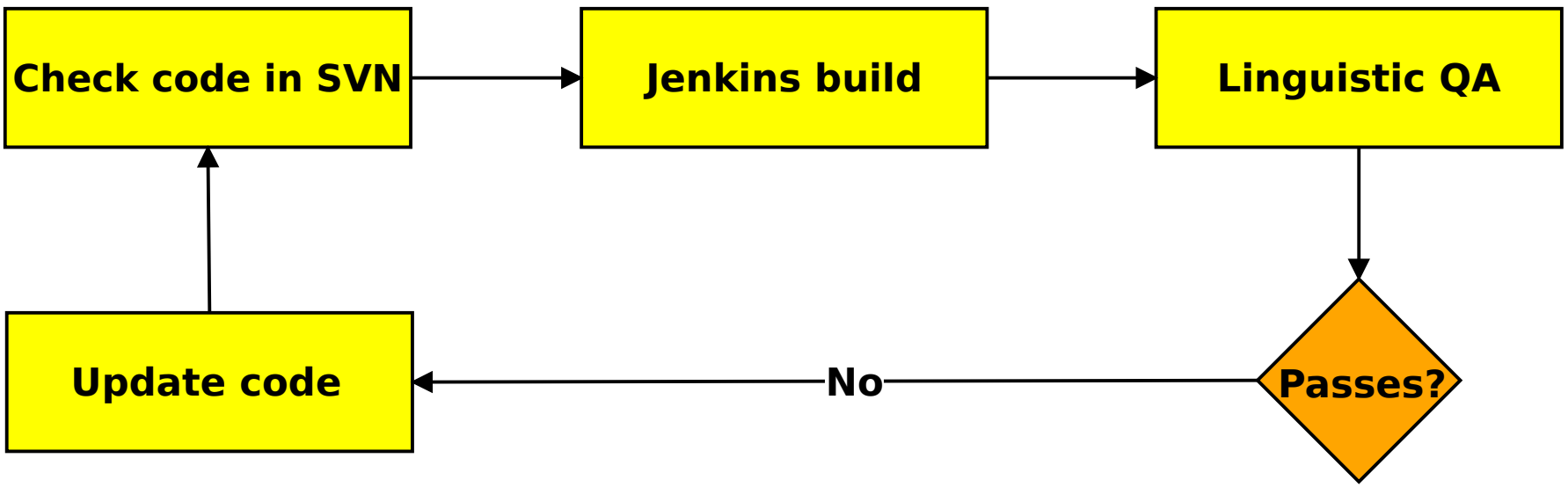
Data Conversion
Build Workflow

OXFORD UNIVERSITY PRESS

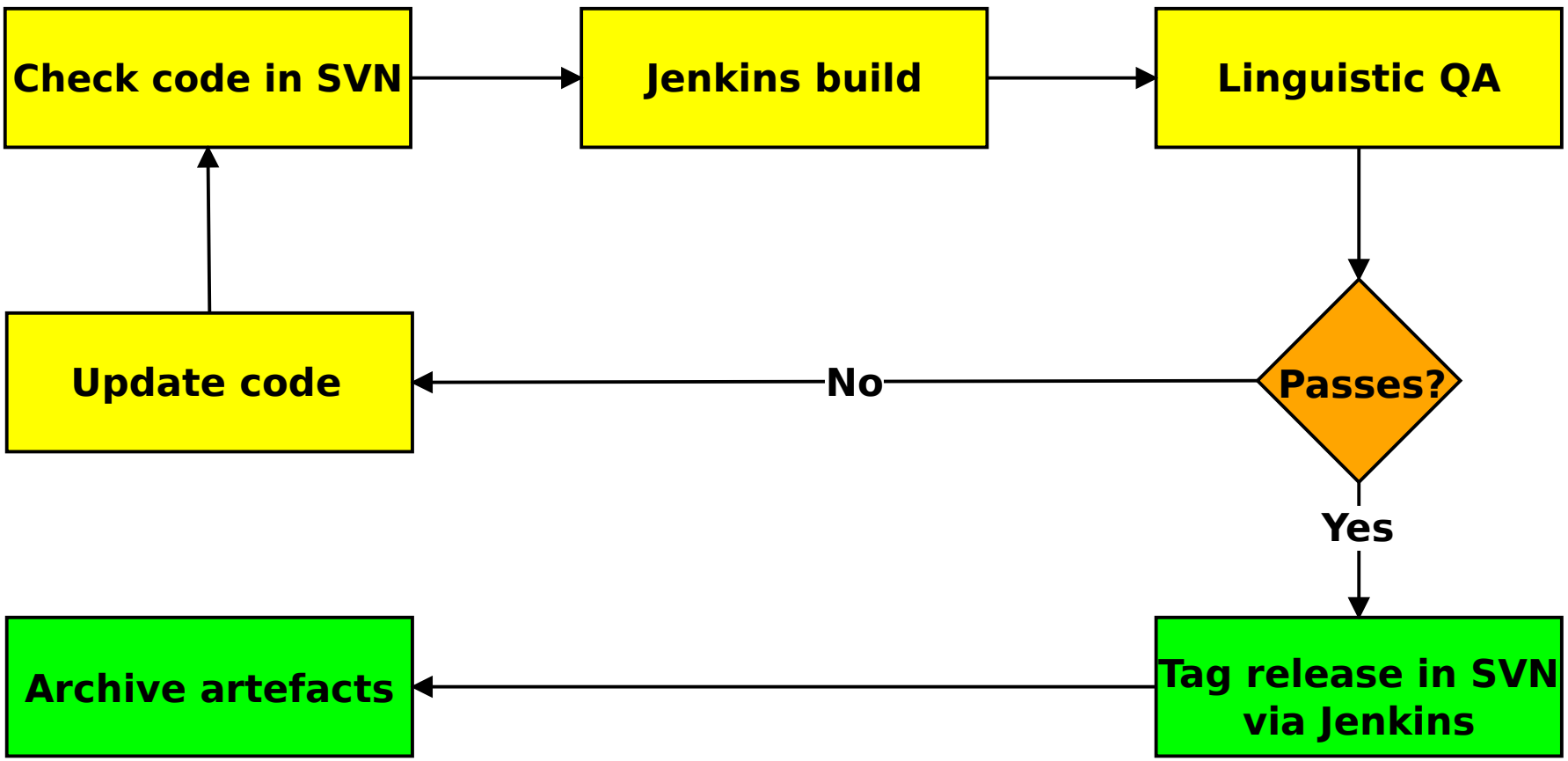


Data Conversion
Build Workflow

Data Conversion
Build Workflow



Data Conversion
Build Workflow



Results & Discussion
Source data

OXFORD UNIVERSITY PRESS

A sense of 'malaúva' from a monolingual Spanish dictionary

```
<ACEPCIO ACEP="2">
  <AREA-GEO>Esp</AREA-GEO>
  <NIVELL>coloquial</NIVELL>
  <SIGNIFICAT>Persona que tiene mal carácter o mala
      intención.</SIGNIFICAT>
  <SINONIM>malaleche.</SINONIM>
</ACEPCIO>
```

Results & Discussion
OUP XML

OXFORD UNIVERSITY PRESS

Print-focused DTD

```
<se2 num="2">
  <lg>
    <ge>Esp</ge>
    <reg>coloquial</reg>
  </lg>
  <msDict type="core">
    <df>Persona que
        tiene mal
        carácter o mala
        intención.</df>
    <syn>malaleche</syn>
  </msDict>
</se2>
```

Results & Discussion
OUP XML

OXFORD UNIVERSITY PRESS

Print-focused DTD

```
<se2 num="2">
  <lg>
    <ge>Esp</ge>
    <reg>coloquial</reg>
  </lg>
  <msDict type="core">
    <df>Persona que
        tiene mal
        carácter o mala
        intención.</df>
    <syn>malaleche</syn>
  </msDict>
</se2>
```

New Lexical XSD

```
<sense register="informal"
    region="ES">
  <definitions>
    <definition>
      <text>Persona que tiene
          mal carácter o mala
          intención</text>
    </definition>
  </definitions>
  <synonyms>
    <synonym>malaleche</
        synonym>
  </synonyms>
</sense>
```

Results & Discussion
OUP XML

OXFORD UNIVERSITY PRESS

Print-focused DTD

```
<se2 num="2">
  <lg>
    <ge>Esp</ge>
    <reg>coloquial</
       reg>
  </lg>
  <msDict type="core">
    <df>Persona que
       tiene mal
       carácter o mala
       intención.</df>
    <syn>malaleche</syn>
  </msDict>
</se2>
```

New Lexical XSD

```
<sense
register="informal"
region="ES">
  <definitions>
    <definition>
      <text>Persona que tiene
          mal carácter o mala
          intención</text>
    </definition>
  </definitions>
  <synonyms>
    <synonym>malaleche</
       synonym>
  </synonyms>
</sense>
```

Next steps
Scale It Up

OXFORD
UNIVERSITY PRESS

- Consolidate data in an **XML database**
- Build an **RDF layer** on top of the XML database
- Leverage **Semantic Web** to enhance our data

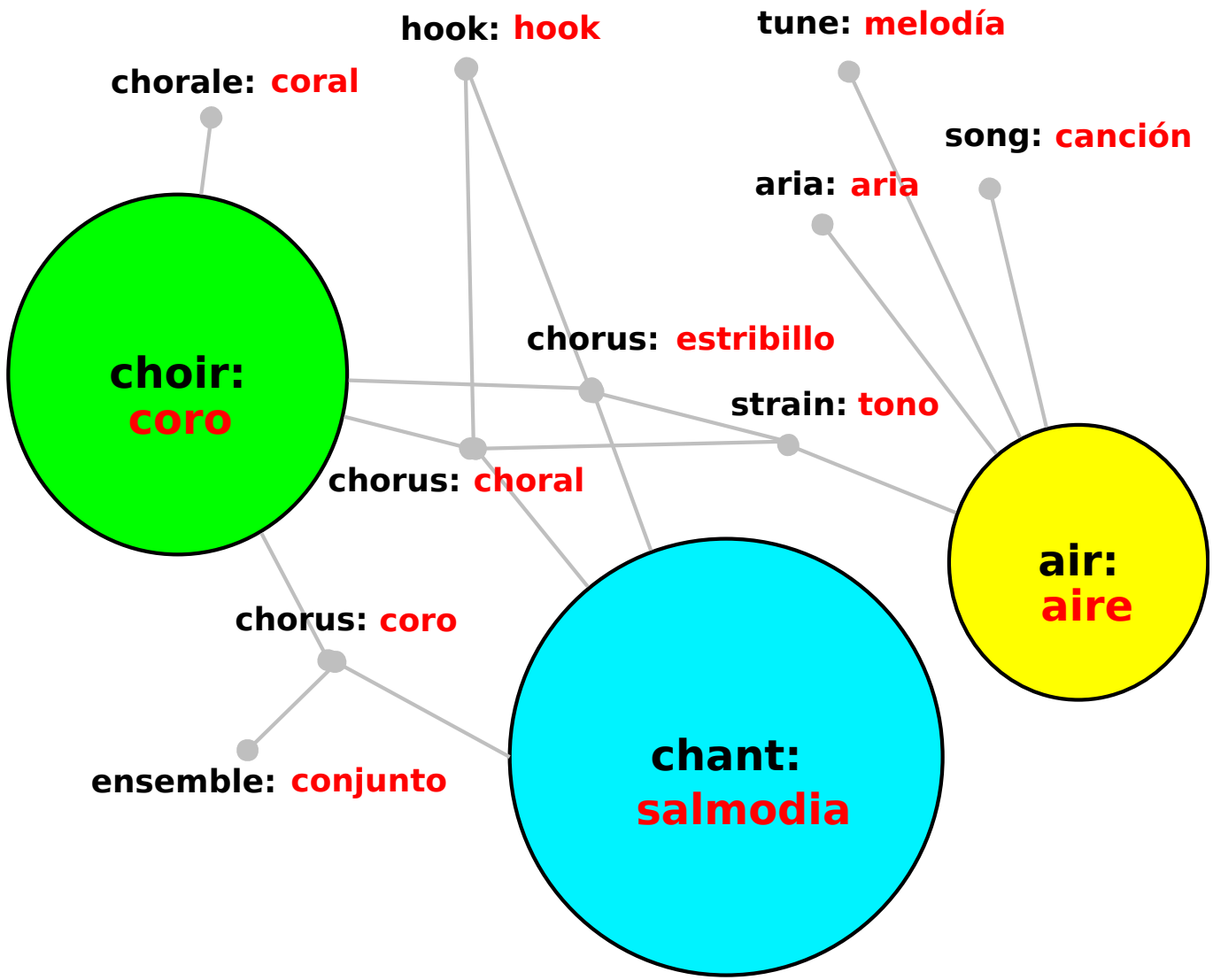
Next Steps
Prototype RDF/XML

OXFORD UNIVERSITY PRESS

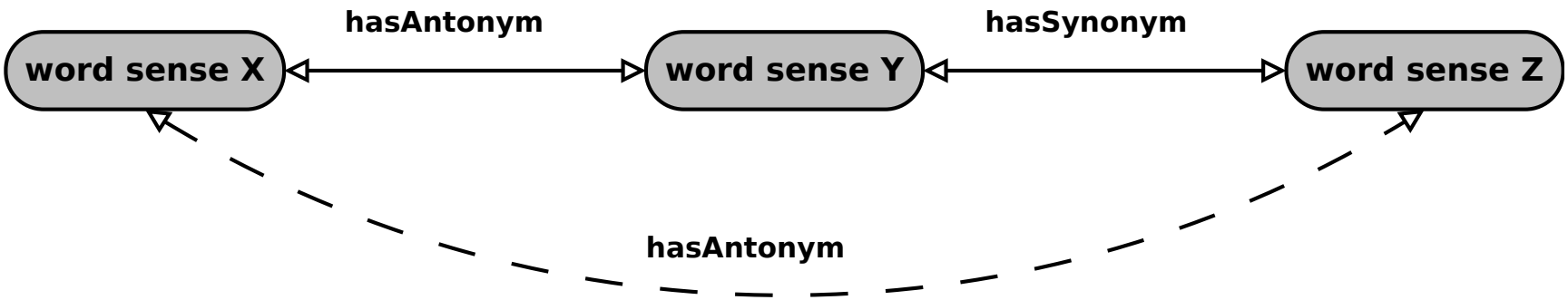
```
<Sense rdf:about="sense:es_noun_malauva_se_2">
  <isDescribedBy
    rdf:resource="
        definition:es_noun_malauva_se_2_def_1"/>
  <hasRegister rdf:resource="register:informal"
     />
  <hasRegion rdf:resource="region:ES"/>
  <hasSynonym rdf:resource="lemma:a5e644"/>
</Sense>
<StandardDefinition
  rdf:about="definition:es_noun_malauva_se_2_def_1">
  <rdfs:label xml:lang="es">Persona que tiene mal
   carácter o mala intención</rdfs:label>
</StandardDefinition>
```

RDF Data extraction

Musical terms in English & Spanish



Inference mechanism



Summary

OXFORD
UNIVERSITY PRESS

- **Overall project requirements**
  - Moving from products to platforms and services
  - Supporting current business needs while innovating
  - Adapting in nimble ways to fast changing market requirements
  - Focusing on time and cost efficiency
- **Data model**
  - Content driven
  - Machine interpretable
  - Modular
  - Evolvable/adaptable
- **Conversion process**
  - Highly automated
  - Modular
  - Scalable

OXFORD
UNIVERSITY PRESS

Thank you for your attention!
Any questions?

Matt Kohl: matt.kohl@oup.com
Sandro Cirulli: sandro.cirulli@oup.com