Semantic Roundtripping for internal Processes and cross-company Data Integration

Cornelsen

Cornelsen in numbers

The publisher Cornelsen is one of the leading providers of educational media in German-speaking countries.

More than 1300 employees work at our head office in Berlin and seven other locations.

1946

Publishing house founded
We have been active in education for more than 70 years

1300

Employees
at Cornelsen Verlag and at the Franz Cornelsen education holding company

22

Nationalities
getting on fantastically well at our offices

5.000

Authors and editors
work with us on inspiring education solutions

53

Professional titles
Under one roof – in addition to publishers, there are scrum masters, business analysts and developers

80

per cent
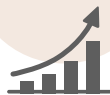of all German teachers use Cornelsen products in their work

1:1

Half
of all managers are female

19.000

Partners
are part of our global network from booksellers to the Goethe Institute

23.000

Titles
And the publishing programme is growing from year to year

4,8 Mio.

Textbooks
sold including access codes for e-books

10.000

teaching materials
Available for download

7,8 Mio.

Every pupil
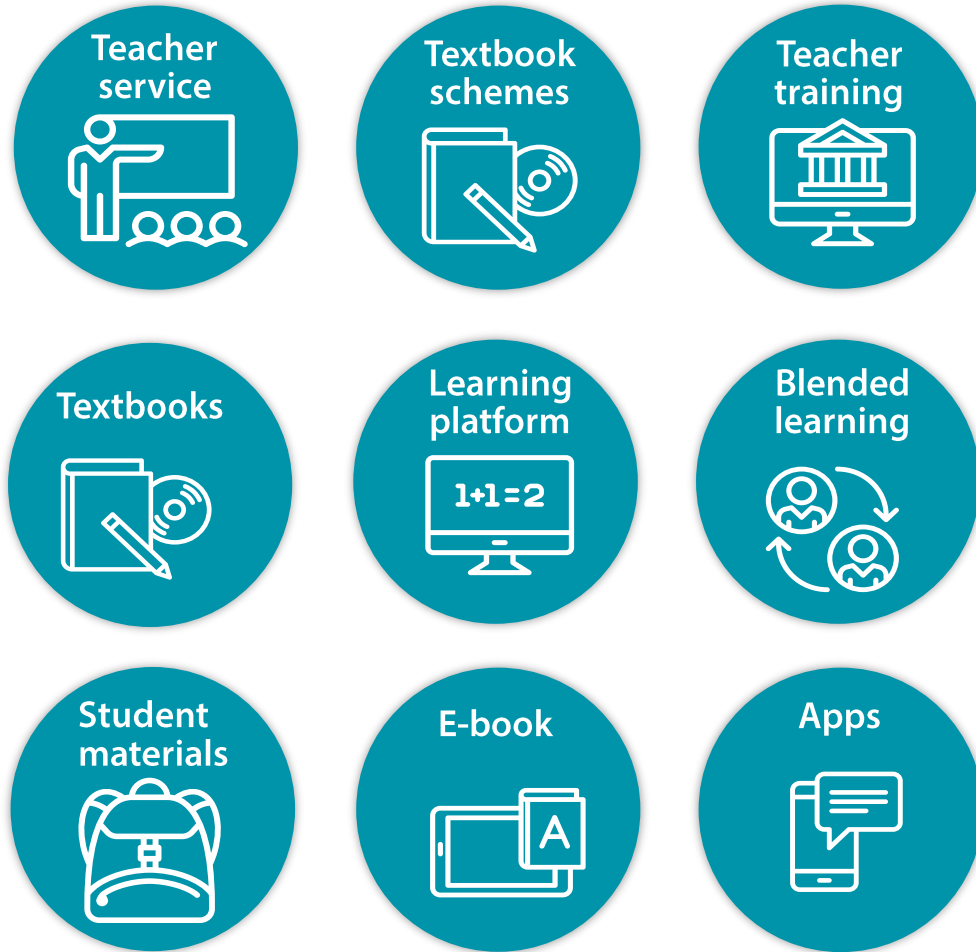in Germany owns at least one Cornelsen workbook

Teachers and pupils trust in our brands

Teachers and pupils trust us because Cornelsen's wide-ranging learning methods and learning media are developed in close cooperation with both parties.

The range is shaped by everyday school life, relieving the pressure on teachers and encouraging effective and lively lessons.

Welcome to the world of learning

Teacher service

Textbook schemes

Teacher training

Textbooks

Learning platform

1+1=2

Blended learning

Student materials

E-book

A

Apps

The publishing programme comprises more than 23,000 titles for all subjects, school types and German states – from early years education to teaching and learning systems for secondary schools and vocational schools, and educational media for adult education and specialist pedagogical literature.

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Goals of our project

- A great variety of existing industries are transformed by semantic technologies
- In these industries, many processes are based on XML
- XML is not just XML!
  - Each industry uses (and sometimes develops) dozens or even hundreds of XML formats
- XML usage and interplay between formats is always embedded in (company internal) processes
- Basic hypothesis: Understanding the role and interplay of XML formats in selected processes will lead to new business opportunities for semantic technologies
  - Bring added value to existing processes
  - Avoid the need to change processes
  - Allow to do semantic processing without breaking formats & processes
- Goal of this project: validate this hypothesis in dedicated use cases

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Project Partners
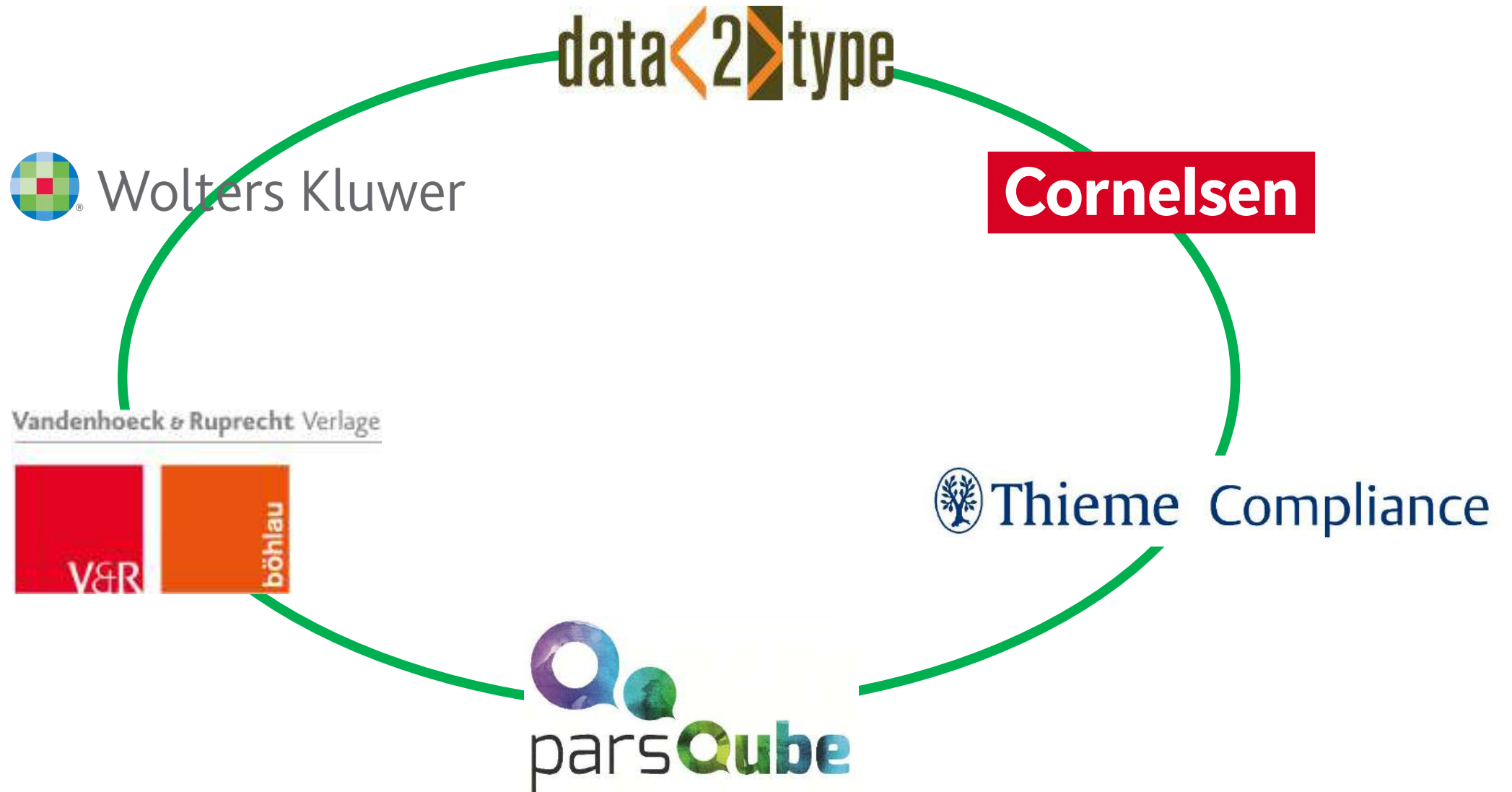
Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Wolters Kluwer Use Case: Automatic content enrichment

Task

- Detect simple entities like names and simple content structures like legal paragraphs in unstructured text
- Detect complex content structures like legal facts in court decisions

- Feed the detected data back into existing XML files or as RDF in the CMS triple store

Goal

- Improve the existing and very efficient XML workflow with additional enrichment capabilities using ML and NLP, so that the process is further simplified, faster and therefore also more cost effective

- This will help to broaden the applicability of semantic web technologies within Wolters Kluwer

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Cornelsen use case: automatic tagging

Birds eye view on educational publishing

```
┌─────────────┐                    ┌─────────────┐
│   Content   │───────────────────▶│ Publication │
│   creation  │                    │             │
└─────────────┘                    └─────────────┘
```
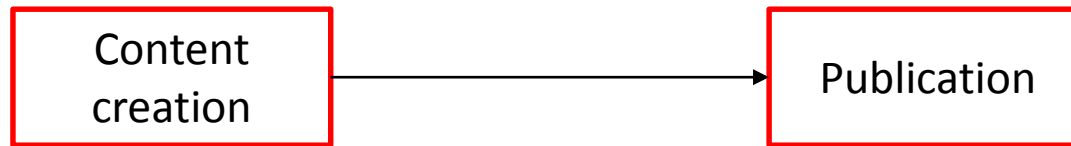
Cornelsen use case: automatic tagging

Reality – a short list of *example* processes involved
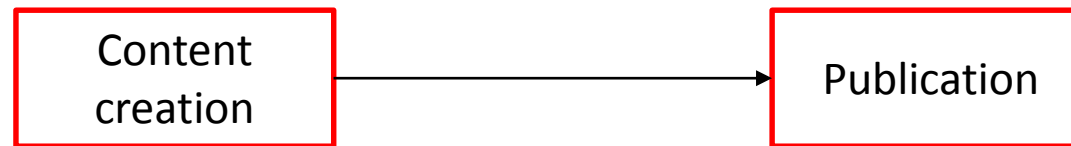
| Content creation | → | Publication |

- Analysis of existing content
- Content model creation
- Content model re-use

- Content transformation
- Transformation output quality assurance (manual + automatic)
- Re-transformation

- Editing environment parametrization
- Content Editing
- Output preview
- Quality assurance

- Digital rights management verification
- Pre-publication
- Product testing
- Actual publication

Cornelsen use case: automatic tagging

Automatic tagging: enrich content with links to internal or public data sources
- Potentially relevant for **several processes**
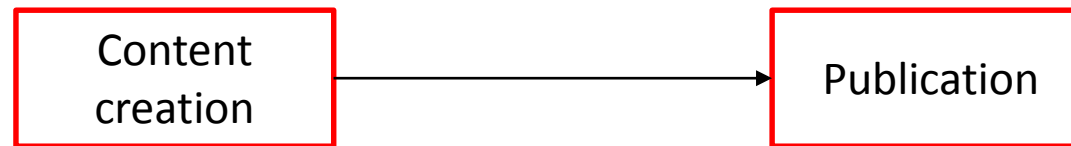
```
┌─────────────┐                          ┌─────────────┐
│ Content     │ ──────────────────────▶  │ Publication │
│ creation    │                          │             │
└─────────────┘                          └─────────────┘
```

- Analysis of existing content
- **Content model creation**
- Content model re-use

- **Content transformation**
- **Transformation output quality assurance** (manual + automatic)
- **Re-transformation**

- **Editing environment parametrization**
- **Content Editing**
- **Output preview**
- **Quality assurance**

- Digital rights management verification
- Pre-publication
- **Product testing**
- Actual publication

Cornelsen use case: automatic tagging

- Semantic technologies without XML integration: several processes need to be changed
  - Content model creation, content transformation, transformation output quality assurance, re-transformation, editing environment parametrization, content editing, output preview, quality assurance, product testing
- Cornelsen uses the format
  - DITA
  - The Learning & Training Content Specialization
  - With Cornelsen specific adaptations
- Main goal of this use case: integrate automatic tagging into DITA L&T CV processes, while avoiding large adaptations to these processes

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Demo: combining XML and RDF processing

Demo: combining XML and RDF processing

Demo: combining XML and RDF processing

Technologies

We work exclusively with XML technologies and use standards such as XSLT, Schematron, XPath, OOXML (Office Open XML).

_rels	10.09.2018 16:56	Dateiordner	
customXml	10.09.2018 16:56	Dateiordner	
docProps	10.09.2018 16:56	Dateiordner	
word	10.09.2018 16:56	Dateiordner	
[Content_Types].xml		XML Document	3 KB

Demo: combining XML and RDF processing

Demo: combining XML and RDF processing

What are the advantages of such an approach?

- The document is created without an API.
- No security problems.
- Future-proof for decades as long as Docx is supported.
- Batch-compatible.
- Can be used for thousands of documents.

Demo: combining XML and RDF processing

What other formats can we integrate in this way?

- Excel
- PowerPoint
- InDesign
- SVG
- EPub
- HTML
- …

The same technologies can also be used to transform DocX into XML, for example, and then move in the direction of RDF.

Demo: combining XML and RDF processing

If you want to test these techniques, you are welcome to try our free online tool data2check.

www.data2check.de

Agenda

- Goals of our project
- Project partners and their motivations
- Wolters Kluwer use case
- Cornelsen use case
- Demo: combining XML and RDF processing
- Conclusions and next steps

Conclusion

- Raise awareness of both Semantic Web and XML Community on mutual benefits
- Challenges for all partners are very similar, although they do different business. Therefore we think that we have a huge scalability potential here

- Existing solutions in both worlds can already help a lot, unfortunately people just do not know enough of each other (low hanging fruit)
- Knowledge engineering and explicit knowledge representation in machine readable form are key for smart content applications. XML and RDF/SKOS are perfect formats to represent this core assets

- We will work on these issues first half of 2019 and report back our results at next SEMANTiCS. Anybody interested? Please talk to us

Contact

Felix Sasaki
Cornelsen Verlag, Publishing Operations
Content Architect
Tel: +49 30 897 85-8398
Email: felix.sasaki@cornelsen.de

Cornelsen Verlag GmbH
Mecklenburgische Straße 53
14197 Berlin

cornelsen.de

Manuel Montero Pineda
data2type GmbH
Managing Director, data2type GmbH
Tel: +49 6221 7391264
Email: montero@data2type.de

data2type GmbH
Wieblinger Weg 92a
69123 Heidelberg

data2type.de

Christian Dirschl
Chief Content Architect
Innovation & UX
Tel: +49 173 93 15 655
Email: Christian.Dirschl@wolterskluwer.com

Wolters Kluwer Deutschland GmbH
Freisinger Strasse 3
85716 Unterschleißheim

wolterskluwer.com