BERLIN 2012 CONFERENCE

17th-19th October

Document Profiling

Dr. Klaus-Peter Eckert, Dr. Stephan Gauch
Fraunhofer FOKUS, Berlin

Gefördert durch:
Bundesministerium für Wirtschaft und Technologie
aufgrund eines Beschlusses des Deutschen Bundestages

Fraunhofer FOKUS

LibreOffice

Agenda

- The context
  - About Fraunhofer FOKUS
  - Document profiling in ODF
- The past
  - Document interoperability in ISO/IEC TR29166
  - Document interoperability in the BMWi “Interop project”
  - Document profiling in the BMWi “Transdok project”
- The results
  - Idea – theoretical concepts
  - Tools – practical support
  - Evaluation – some math
- Conclusion

The Context

Gefördert durch:

Bundesministerium
für Wirtschaft
und Technologie

aufgrund eines Beschlusses
des Deutschen Bundestages

About Fraunhofer

Fraunhofer FOKUS

The Fraunhofer Gesellschaft is Europe‘s largest organization for applied research

- Fraunhofer develops products and processes through to technical or commercial maturity
- Individual solutions are elaborated in direct contact with the customers
- The Fraunhofer Gesellschaft maintains
  - 60 self-contained Fraunhofer Institutes throughout Germany
  - with a staff of 20, 000 scientists and engineers
  - 1.8 billion Euro annual budget
- 70% of funding are raised through innovative development projects, license fees and contract research
- Sub-companies and representative offices all over the world

About FOKUS

Fraunhofer institute for Open CommUnication Systems

- Telephony: FOKUS developed SIP
- Web2.0 / Web / Telco Convergence
- Model-driven Engineering
- IPTV & Rich Media
- Future Internet
- Smart Mobility
- Smart Energy
- **eGovernment: One stop shopping**
- Test automation: Invention of TTCN-3
- IMS – Next Generation Networks Platforms & Services
- Electronic Safety and Security Systems
- eHealth
- founded 1988 (HMI, GMD, FhG)
- more than 500 employees

Smart Cities

The FOKUS Interoperability Labs

Fraunhofer FOKUS

ICT solution provider

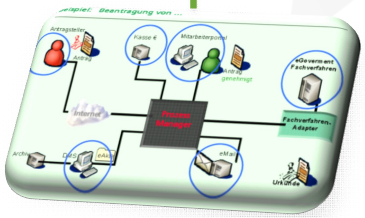
ICT customer
Public administrations

Workshop

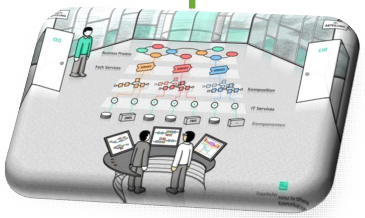
Showcase

Competence node

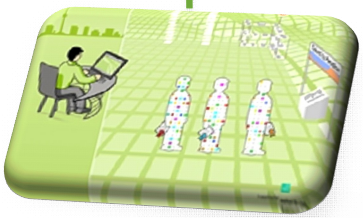
Interoperability-Labs



eGovernment IOP-Lab

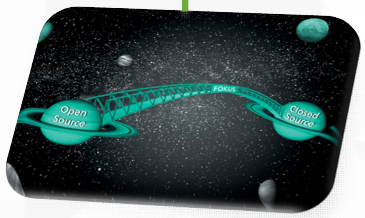


Cloud/SOA IOP-Lab



Secure eIdentity IOP-Lab

Document IOP-Lab



Open/Closed Source IOP-Lab

Document IOP-Lab: The Big Picture

Fraunhofer FOKUS

OOXML semantic validation and profiling

OOXML portability and interoperability

Office

OOXML archiving PDF/A

ODF-OOXML ISO-TR 29166

ODF-OOXML translation

Adobe PDF

ODF archiving PDF/A

ODF and OOXML Test Library

ODF semantic validation and profiling

ODF portability and interoperability

Microsoft

ISO DIN

Gefördert durch:

Bundesministerium für Wirtschaft und Technologie

aufgrund eines Beschlusses des Deutschen Bundestages

Portal: http://www.interoperability-center.com/en/dokumenteninterop-lab

Document Interoperability
Where the Shoe Pinches

Fraunhofer FOKUS

- The joint processing of documents such as text documents, presentations and spreadsheets fails due to the incompatibility of the used tools
  - Tools process documents using specific interpretations
    - Content may get lost
    - Documents may be represented in different ways
  - Document formats cannot be converted without loosing some structure and content
- Received documents and data do not follow the rules of the recipient, forwarded documents and data do not follow the rules of the addressee
  - Paperless processes are only supported to a limited extend
- Archived documents may not be reopened and processed

Conformity and Interoperability Discussion in OASIS ODF OIC

Fraunhofer FOKUS

- The OASIS OIC TC (OASIS Open Document Format Interoperability and Conformance) helps implementors create applications that conform to the OpenDocument Format (ODF) OASIS Standard. ODF defines a genuinely open XML file format for office productivity applications, including text, spreadsheets, charts, graphs, presentations, and databases. The OIC TC works to ensure that the growing number of ODF-compliant applications are able to interoperate and conform to the standard.
- Deliverables
  - Initial report on the state of ODF conformance and interoperability
    - State of interoperability V1.0; CS 01, December 2010
  - Report on the best practices on profiles and recommendations on possible ODF-related profiles
    - ODF 1.1 Interoperability Profile; CD 03; June 2010
  - A conformity assessment methodology specification, detailing how each provision and recommendation in the ODF standard may be tested for conformance

Excerpt from State of ODF Interoperability

Fraunhofer FOKUS

- According to ISO/IEC 2382-01, “Information Technology Vocabulary, Fundamental Terms”, interoperability is the capability to
  - communicate,
  - execute programs, or
  - transfer data

  among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.

- From the perspective of ODF,
  - the document is the data which is transferred, and
  - the functional units are the software applications which create, edit, view and manipulate these documents.
- Interoperability is high if the document can be successfully transferred among such applications, without the user needing to be concerned with the unique characteristics of each application.

Excerpt from State of ODF Interoperability (cont.)

Fraunhofer FOKUS

- Since the capabilities of ODF applications extend beyond the common desktop editors, and include other product categories such as web-based editors, mobile device editors, document convertors, content repositories, search engines, and other document-aware applications, interoperability will mean different things to users of these different applications. However, to one degree or another, interoperability consists of meeting user expectations regarding one or more of the following qualities when transferring documents:
  - The visual appearance of the document at various levels, e.g. glyph, run, line, block, page
  - The structure of the document as revealed when the user attempts to edit the document, e.g., headers, paragraphs, lists, tables
  - The behaviors and capabilities of internal and external links and references
  - The behaviors and capabilities of embedded images, media and other objects
  - The preservation of document metadata
  - The preservation of document extensions
  - The integrity of digital signatures and other protection mechaniss.
  - The runtime behaviors manifest from scripts, macros and other forms of executable logic

Conformity in ODF
Excerpt from IS 26300

Fraunhofer FOKUS

- Documents that conform to the OpenDocument specification may contain elements and attributes not specified within the OpenDocument schema. Such elements and attributes must not be part of a namespace that is defined within this specification and are called foreign elements and attributes.

- Conforming applications either shall read documents that are valid against the OpenDocument schema if all foreign elements and attributes are removed before validation takes place, or shall write documents that are valid against the OpenDocument schema if all foreign elements and attributes are removed before validation takes place.

- Conforming applications that read and write documents may preserve foreign elements and attributes. In addition to this, conforming applications should preserve meta information and the content of styles.

Conformity in ODF
Excerpt from IS 26300

- Foreign elements may have an office:process-content attribute attached that has the value true or false. If the attribute's value is true, or if the attribute does not exist, the element's content should be processed by conforming applications. Otherwise conforming applications should not process the element's content, but may only preserve its content. If the element's content should be processed, the document itself shall be valid against the OpenDocument schema if the unknown element is replaced with its content only.

- Conforming applications shall read documents containing processing instructions and should preserve them. There are no rules regarding the elements and attributes that actually have to be supported by conforming applications, except that applications should not use foreign elements and attributes for features by the OpenDocument schema.

The Past

- ISO/IEC SC34 WG5 TR29166, September 2011, Busan
  - Guidelines for translation between ISO/IEC 26300 and ISO/IEC 29500 document formats
    Approved as TR (Klaus-Peter Eckert, Ed.)
- ODF Plugfest, June 2011, Berlin
  - Utilization of Document Test Libraries Supporting the Interoperability of Office Applications
    Lessons learned (Klaus-Peter Eckert)
  - Translatability of Document Formats
    Feature Driven Profiling of Open Standards for Office Applications (Björn Kirchhoff)

Interoperability, Conformity, Profiles Technical Definitions

Fraunhofer FOKUS

- Interoperability and conformity: ISO JTC1 Directives, Annex I (IT standards)
  - Standards designed to facilitate interoperability need to specify clearly and unambiguously the conformity requirements that are essential to achieve the interoperability. ..Verification of conformity to those standards should then give a high degree of confidence in the interoperability of IT systems using those standards. However, the confidence in interoperability given by conformity to one or more standards is not always sufficient and there may be need to use an interoperability assessment methodology in demonstrating interoperability between two or more IT systems in practice…An assessment methodology for interoperability may include the specification of some or all of the following: terminology, basic concepts, requirements and guidance concerning test methods, the appropriate depth of testing, test specification and means of testing, and requirements and guidance concerning the operation of assessment services and the presentation of results. In technical areas where there is a conformity assessment methodology and an interoperability assessment methodology, the relationship between them must be specified.
- Profile: ISO concept database - ISO 14772 (Virtual Reality Modeling Language)
  - A named collection of criteria for functionality and conformance that defines an implementable subset of the standard

Conformity and Interoperability Discussion within ISO SC34

Fraunhofer FOKUS

- Conforming
  - Obeys the provisions of a specification
  - Conformance tests enable a better correspondence between what the standard says and what exists in reality
- Valid
  - Of an XML document, that it obeys a schema
- Interoperable
  - A property of systems that interoperate
- Portable
  - A property of data/document that may be used interoperable

Document Properties
ISO/IEC SC34 WG5 TR 29166

Document models

ODF "State of Interoperability"

TR29166 Document features

Improved document interoperability

Feature based profiles

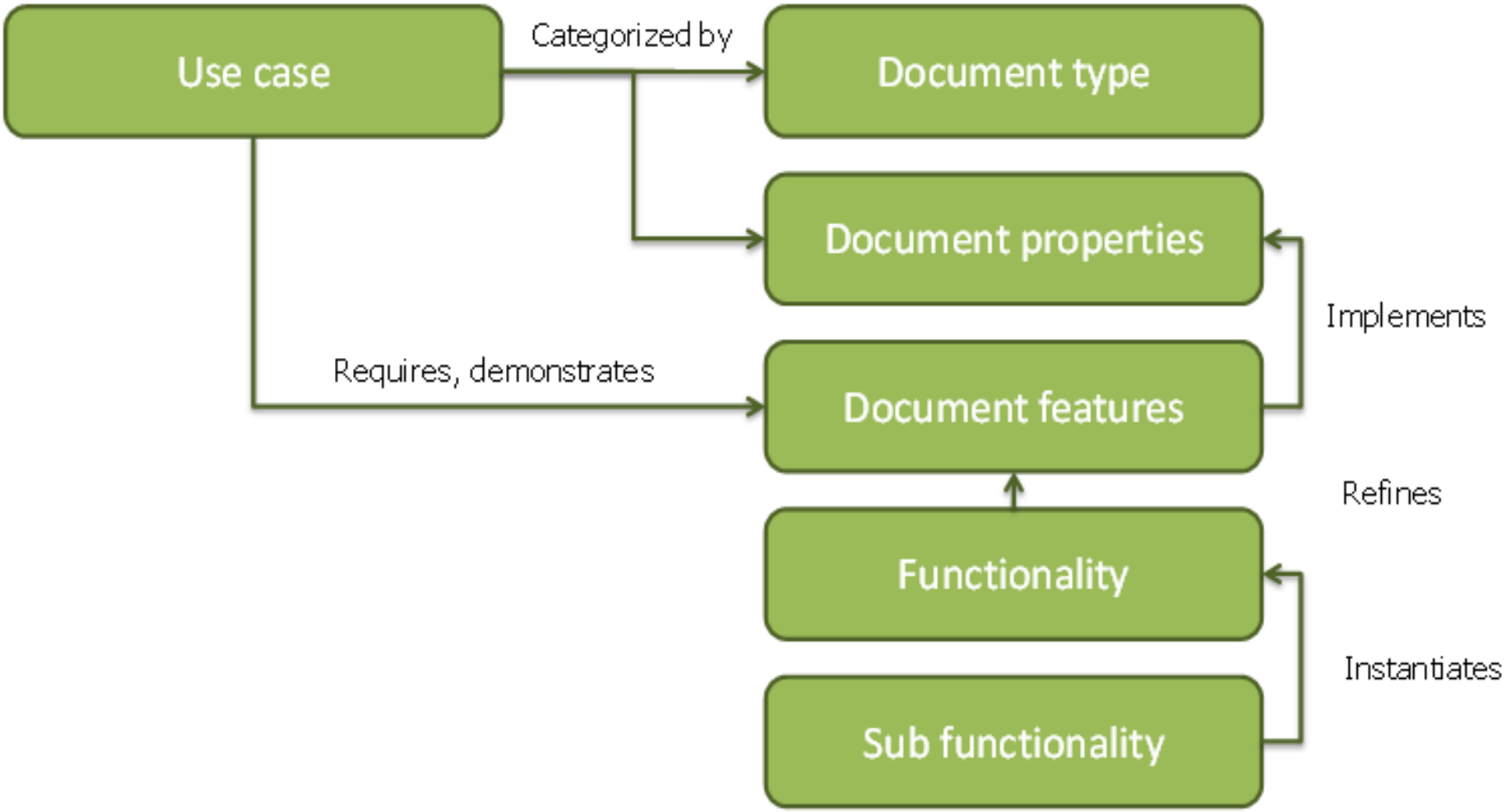
Interoperability metrics

Document Properties ISO/IEC SC34 WG5

Fraunhofer FOKUS

- Presentation instructions include layout and presentation related information such as fonts, spacing, margins, color, and animation in office documents.
- Document content covers all properties of content such as text, graphics and formulas defined directly by the author of a document.
- Dynamic content covers all aspects of automatically generated content, calculations or form functionalities such as fields, generated tables, or dynamic references.
- Meta data cover all information apart from the core document content. Metadata are used to describe meta information about the document such as generator, version, authors, and to ensure the accessibility of documents, for instance by using certificates.
- Annotation covers all aspects about annotations to a document, change tracking, and collaborative functions.
- Document parts cover all aspects (editing semantics) of structural document properties such as paragraphs, headings, headers, footers, tables, lists, tables, footnotes, indices, and captions.

Document Features
ISO/IEC SC34 WG5 - TR29166



Sample Document Features ISO/IEC SC34 WG5 - TR29166

- Word processing documents
  - Text formatting
  - Paragraph formatting
  - Header and footer
  - Tables
  - Itemization and numeration
  - Metadata
  - Indices
  - Change tracking
  - Forms
  - Formulas
  - ...
- Spreadsheets
  - Formatting
  - Calculation
  - ...
- Presentations
  - Slides
  - Text formatting
  - Master layouts
  - ...
- Common features
  - Alternative representations
  - Color models
  - ...

Sample Document Features ISO/IEC SC34 WG5 - TR29166

- Feature
  Text formatting
  - Functionality
    - Font weight
    - Text borders
    - Whitespaces
    - Capitalization
      - All upper case
      - Small caps
      - All lower caps
    - Text color
      - RGB
      - Background color
      - Color theme
      - Blinking text
      - Text highlighting
    - Complex script support

- Feature
  Text formatting
  - Functionality
    - East Asian text
    - Font selection
      - By name
      - By family
      - By theme
    - Font effects
    - Run/span width
    - Italic text
    - Kerning
    - Text language
    - Spell checking
    - Raised/lowered text
    - Strike-through
    - Underline

ISO/IEC SC34 WG5 - TR29166

Fraunhofer FOKUS

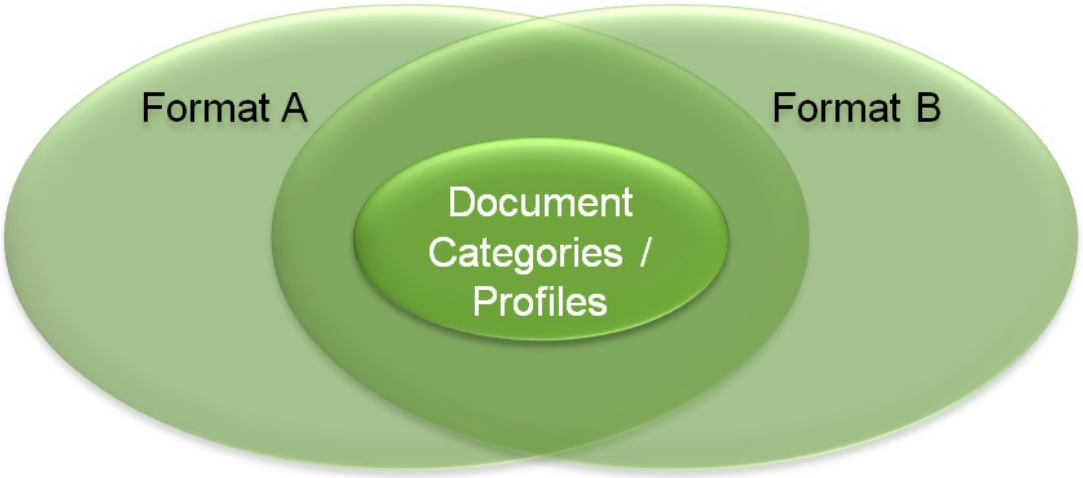
- Guidelines for translation between ISO/IEC 26300 and ISO/IEC 29500 document formats
- ToC
  - Use cases
    - Typical user stories for textprocessing, spreadsheet and presentation docs
  - Features and functionality
    - Comparison of document features ODF and OOXML
  - Samples of typical XML structures
  - Translation complexity between ODF and OOXML
    - Easy (1:n), moderate (n:m) and difficult (impl) translation between both formats
  - Guidelines
    - Conformance tests
    - Interoperability tests
    - Test libraries
    - Profiles and templates
    - Taxonomies, ontologies

Document Categories and Templates Profiling

- Reduce the complexity and features to be used within documents
- Definition of translatable document templates for special purposes

- Improve the translatability of documents
- Additional conformity tests are necessary – membership in a category / to a profile

- ODF profile: improve interoperability
- Categories: improve portability



Sample Document Categories

Document Category	Description
Memo	Documents used to formally present a small amount of information (e.g. an initiative or concept).
Meeting Notes	Documents used to capture the results of a meeting– often informal, created ad-hoc during the meeting itself.
Research Papers / Reports	Documents that are highly technical and conform to a specific layout – common usage of academic features like equations/bibliography.
Articles	Documents that contain carefully formatted content intended for publication in a magazine/journal.
Legal Documents	Documents used in legal proceedings – must conform to a specific layout/form, and are typically built over many revisions by many authors.
Résumés	Short, highly structured documents intended to convey information in a single page. Formatting is used to customize the structured base.
Proposals	Documents containing a customer proposal, compiled from several authors and many sources (Excel, PPT, etc.).
Essays	Documents containing several pages of basic, unformatted text, without specific formatting requirements.
Books	Documents that are used to compile the manuscript of a book in a format ready for publication.
Flyers	Documents used as marketing materials – single page, designed for 6+ foot readability.
Newsletters	Documents containing multiple streams of specially laid out content (typically a fixed layout making use of 2D aspects of the page).
Letters	Documents containing plain, unformatted text, intended for personal communication.
Notes / Lists	Documents that are created on demand to store informal information (e.g. brainstorming data).
Labels	Documents that contain a page of label templates, which are merged with addresses and printed.
Outlines	Documents that contain an outline of a larger document, but little/none of the actual contents.
Structured Document/Form	Documents that contain structured content intended to be populated from data sources (e.g. a database) or form input from the user.

BMWi Project about the Improvement of Interoperability

- Goal of the project:
  - Improve interoperability by the definition of interoperability related
    - Requirements and use cases
    - Methodology (IEEE 829)
    - Test concepts (ETSI)
    - Test labs inc. test libraries
    - Tool support, especially for public procurement
- Application areas:
  - Document interoperability
    - Reuse feature definitions
    - Reuse category/profile definitions
  - Services and processes
  - Identity and access management

The IOP-Assistent

Fraunhofer FOKUS

Navigation

Name

- Kategorien
  - Standards
  - Services
  - IOP-Eigenschaften
    - Prozess-/Diensteigenschaften
    - Sicherheitseigenschaften
    - Dokumenteneigenschaften
      - Präsentationen
      - Tabellenkalkulationen
      - Textdokumente
        - Content, Inhalte
        - Typen und Profile
        - Änderungsverfolgung
        - Annotationen
        - Charts
        - Daten
        - Erweiterungen
        - Absatzformatierung
        - Fließtextformatierung
        - Fonts
        - Formeln
        - Formulare
        - Fuß- und Endnoten
        - Grafiken

Document features

Anforderungskataloge | Anwendungsfälle | Szenarien | Parametrisierungen | Akt

Testdaten in Textdokumente,

Name	Autor
Indizierung OOXML	Klaus-Peter Eckert
Protokoll OOXML	Klaus-Peter Eckert
Protokoll ODF	Klaus-Peter Eckert
Bestellung ODF	Klaus-Peter Eckert
Bestellung OOXML	Klaus-Peter Eckert
Metadaten OOXML	Klaus-Peter Eckert
Metadaten ODF	Klaus-Peter Eckert
Tabellen OOXML	Klaus-Peter Eckert
Tabellen ODF	Klaus-Peter Eckert
Zeilennummern OOXML	Klaus-Peter Eckert
Zeilennummern ODF	Klaus-Peter Eckert
Signatur OOXML (Text)	Klaus-Peter Eckert
Signatur ODF (Text)	Klaus-Peter Eckert
Zugriffssperre OOXML (Text)	Klaus-Peter Eckert
Zugriffssperre ODF (Text)	Klaus-Peter Eckert
Handbuch ODF	Klaus-Peter Eckert
Handbuch OOXML	Klaus-Peter Eckert
Rechtschreibung ODF	Klaus-Peter Eckert
Rechtschreibung OOXML	Klaus-Peter Eckert
Listen ODF	Klaus-Peter Eckert
Listen OOXML	Klaus-Peter Eckert
Kopf- und Fusszeilen ODF	Klaus-Peter Eckert

Document library

The IOP-Assistent

Fraunhofer FOKUS

Navigation

Name

- Kategorien
  - Standards
  - Services
  - IOP-Eigenschaften
    - Prozess-/Diensteigenschaften
    - Sicherheitseigenschaften
    - Dokumenteneigenschaften
      - Präsentationen
      - Tabellenkalkulationen
      - Textdokumente
        - Content, Inhalte
        - Typen und Profile
          - Lebenslauf
          - Protokoll
          - Antrag
          - Bestellung/Rechnung
          - Memo
          - Flyer
          - Brief
          - Handbuch
          - Veröffentlichung

Profiles Categories

Anforderungskataloge | Anwendungsfälle | Szenarien | Parametri

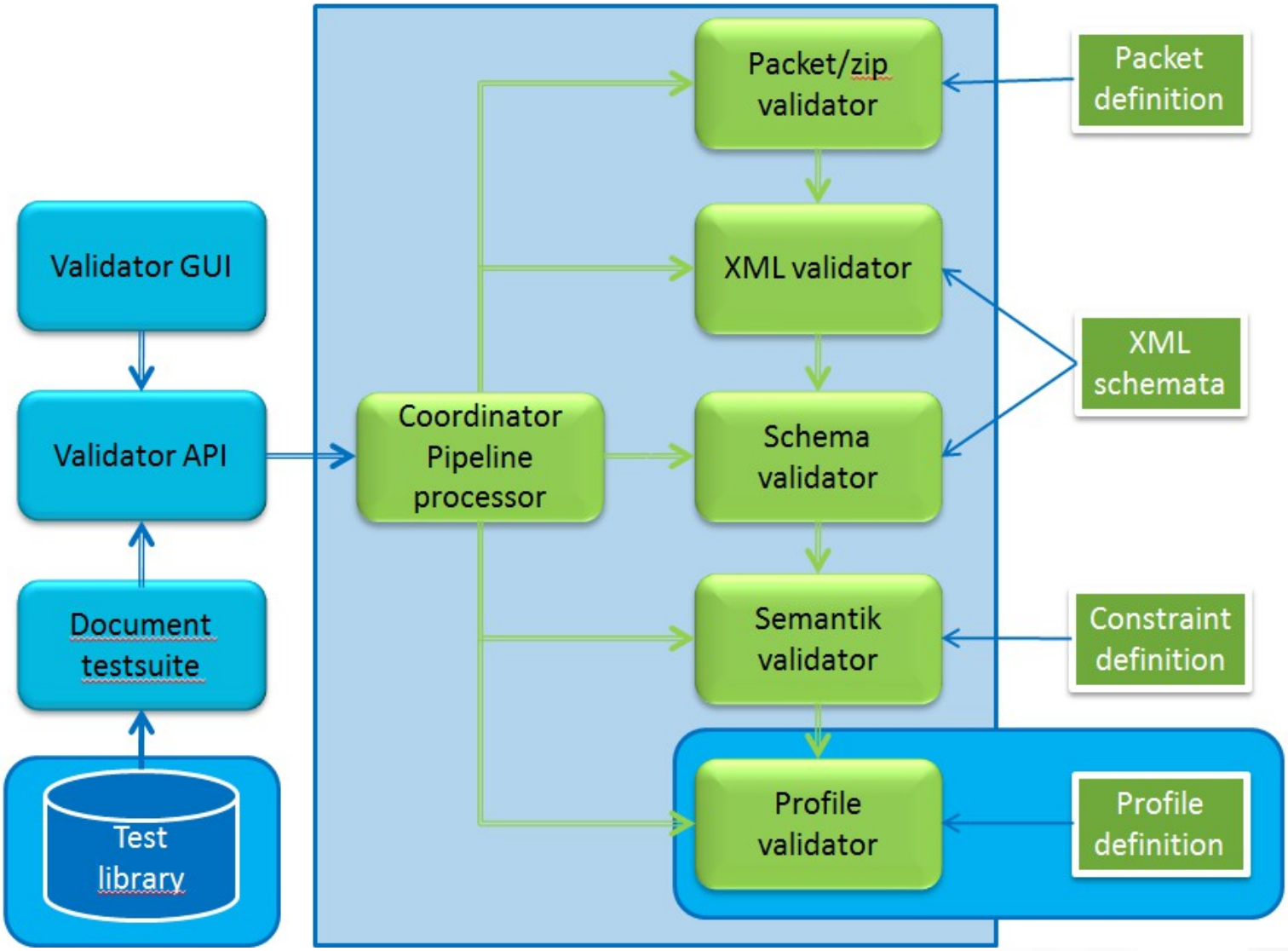
Name	Autor
Textdokument Typ Anträge	Lena-Sophie I
Textdokument Typ Protokolle	Lena-Sophie I
Textdokument Typ Lebensläufe	Lena-Sophie I
Textdokument Typ Veröffentlichungen	Lena-Sophie I
Textdokument Typ Handbuch	Lena-Sophie I
Textdokument Typ Briefe	Lena-Sophie I
Textdokument Typ Bestellungen	Lena-Sophie I
Textdokument Typ Flyer	Lena-Sophie I
Textdokument Typ Memo, Mitschrift	Lena-Sophie I
Textdokument Typ Rechnung	Lena-Sophie I
Templates für Textdokumente	Klaus-Peter E
Textdokument Typ Bestellungen (erweitert)	Lena-Sophie I
Textdokument Typ Briefe (erweitert)	Lena-Sophie I
Textdokument Typ Flyer (erweitert)	Lena-Sophie I
Textdokument Typ Lebensläufe (erweitert)	Lena-Sophie I
Textdokument Typ Memo, Mitschrift (erweitert)	Lena-Sophie I
Textdokument Typ Protokolle (erweitert)	Lena-Sophie I
Textdokument Typ Rechnung (erweitert)	Lena-Sophie I
Textdokument Typ Veröffentlichungen (erweitert)	Lena-Sophie I
Textdokument Typ Handbuch (erweitert)	Lena-Sophie I

Test scenarios

Generation of IEEE 829 conforming test documentation inc. IOP requirements

Document Test & Validation Environment as Part of the Document IOP lab

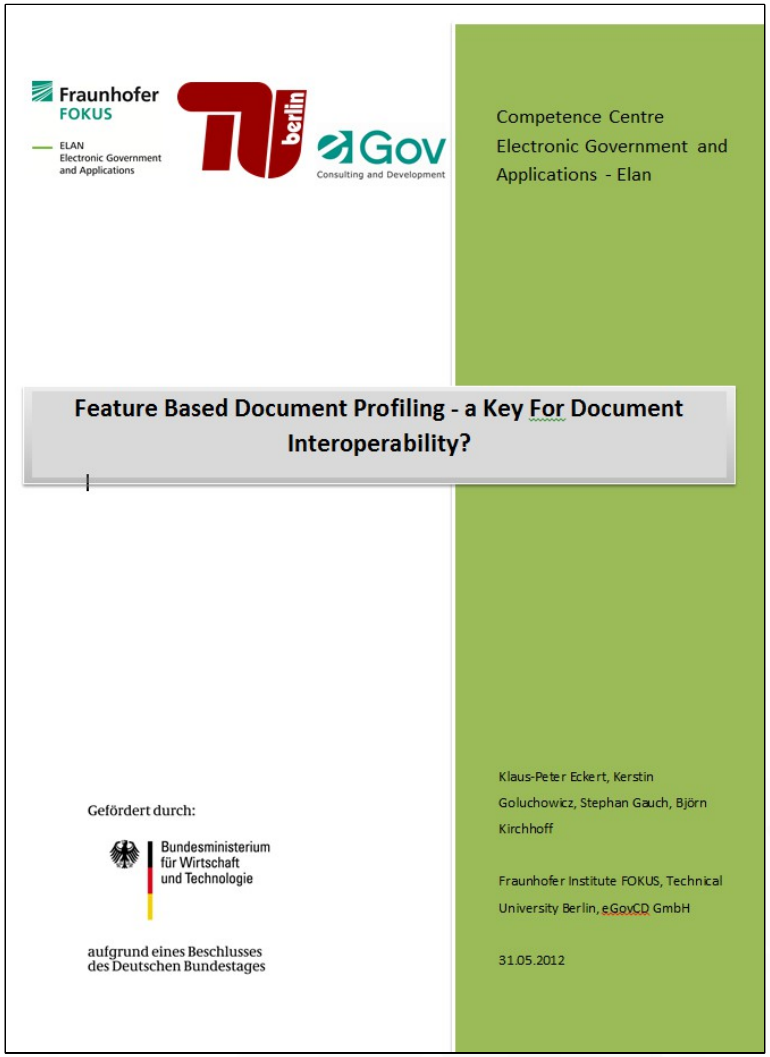
Fraunhofer FOKUS



The Results

- Results of the Transdok project

"Validation and transformation of selected profiles of the document standards ISO/IEC 26300 and ISO/IEC 29500"



Fraunhofer FOKUS

ELAN Electronic Government and Applications

berlin

eGov Consulting and Development

Competence Centre Electronic Government and Applications - Elan

Feature Based Document Profiling - a Key For Document Interoperability?

Gefördert durch:

Bundesministerium für Wirtschaft und Technologie

aufgrund eines Beschlusses des Deutschen Bundestages

Klaus-Peter Eckert, Kerstin Goluchowicz, Stephan Gauch, Björn Kirchhoff

Fraunhofer Institute FOKUS, Technical University Berlin, eGovCD GmbH

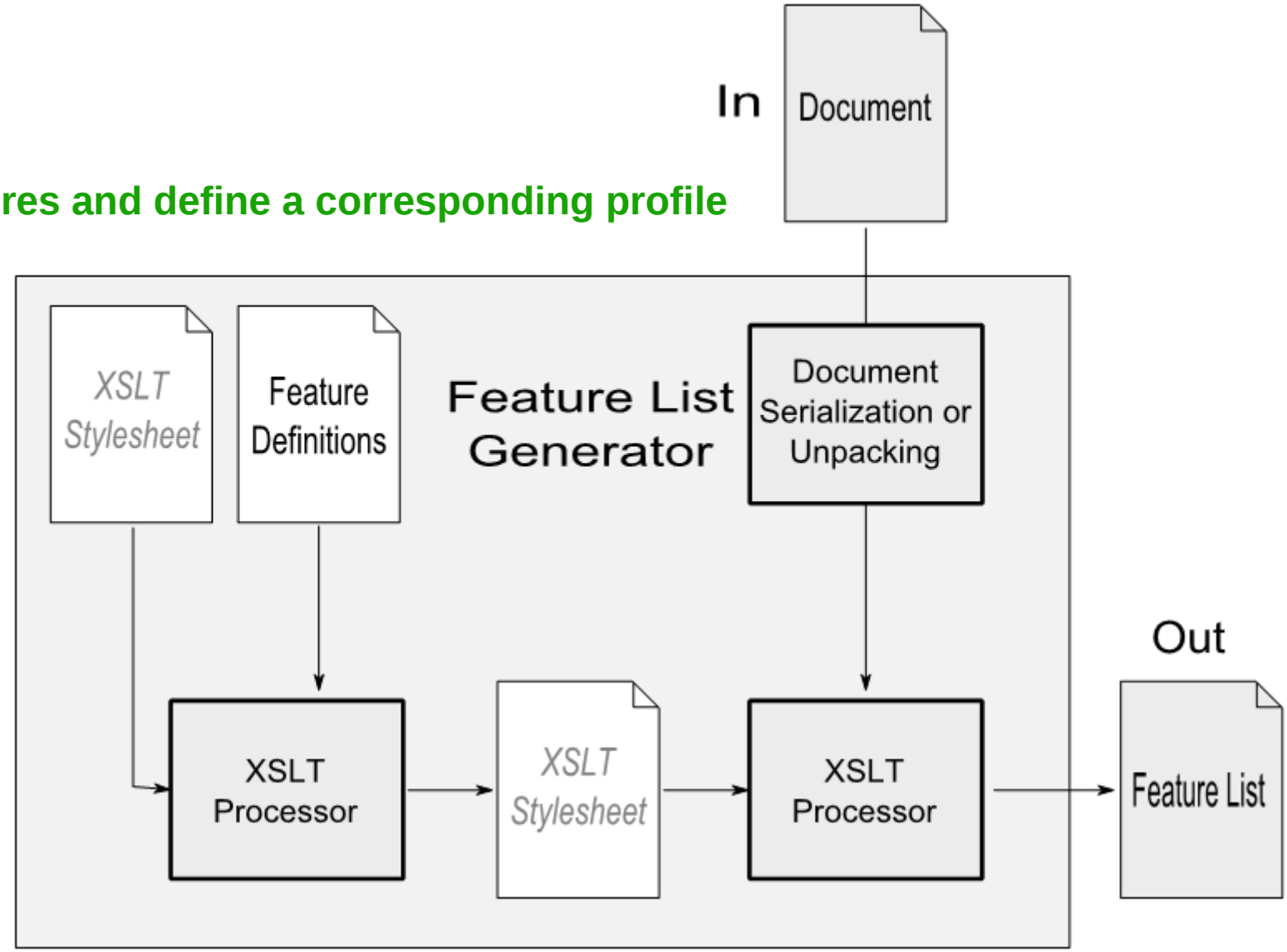
31.05.2012

http://www.interoperability-center.com/de/dokumenten-iop-lab

The Transdok Feature List Generator

Fraunhofer FOKUS

Detect features and define a corresponding profile



The Transdok Feature List Generator
Sample Feature Detection Functions

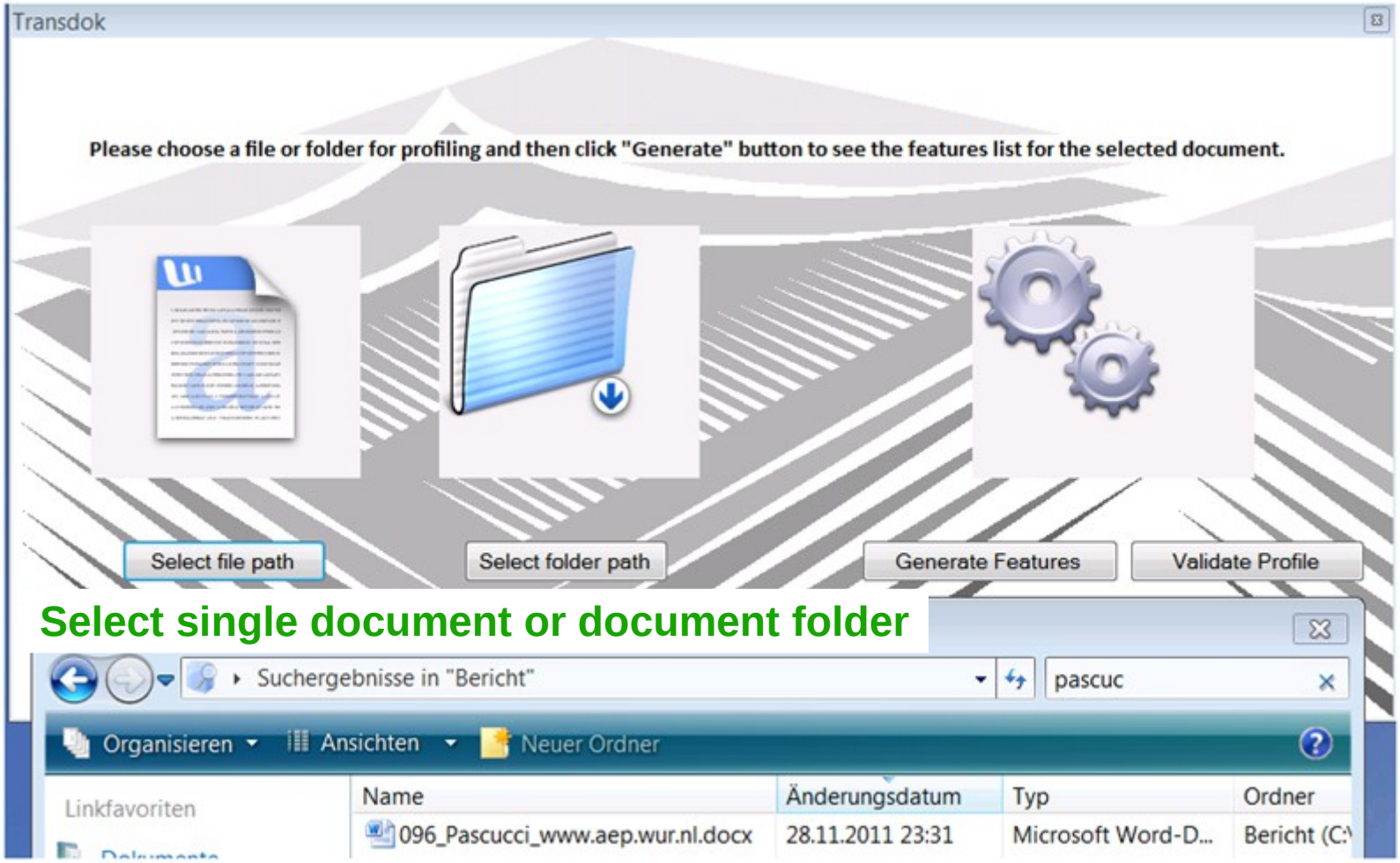
```
<document:feature-category name="Footnotes and endnotes">

    <document:feature name="Footnote">
        <document:standard name="OOXML">
            <document:detectionfunction
                xpath="//ooxml-w:footnoteReference"/>
        </document:standard>
        <document:standard name="ODF">
            <document:detectionfunction
                xpath="//odf-text:note[@odf-text:note-class='footnote']" />
        </document:standard>
    </document:feature>
```

Define feature detection functions for chosen standards utilizing XPath expression

```
    <document:feature name="Endnote">
        <document:standard name="OOXML">
            <document:detectionfunction xpath="//ooxml-w:endnoteReference"/>
        </document:standard>
        <document:standard name="ODF">
            <document:detectionfunction
                xpath="//odf-text:note[@odf-text:note-class='endnote']" />
        </document:standard>
    </document:feature>
```

The Transdok Feature List Generator Tool



Select single document or document folder

Summary Listing from the Feature List Generator

Fraunhofer FOKUS

s list f ... ocuments\elan\transdok\transdokDateie

Document feature

Document functionality

Feature Category	Feature Name	Total files	Relative Total Files	Total usages
Absatzformatierung		761	1,00	502606
Absatzformatierung	Absatzformat - Posi...	57	0,07	57
Absatzformatierung	Absatzformat - Ausri...	1	0,00	2
Absatzformatierung	Absatzformat - Bloc...	501	0,66	44100
Absatzformatierung	Absatzformat - Einz...	719	0,94	132149
Absatzformatierung	Absatzformat - Hinte...	177	0,23	2323
Absatzformatierung	Absatzformat - Initial...	9	0,01	19
Absatzformatierung	Absatzformat - Links...	704	0,92	1616
Absatzformatierung	Absatzformat - Posit...	259	0,34	2064
Absatzformatierung	Absatzformat - Rah...	253	0,33	2584
Absatzformatierung	Absatzformat - Rech...	370	0,49	20911

Back to documents list

CSV export

Profile Definition in the Feature List Generator

Fraunhofer FOKUS

Define document profile based on a statistic analysis of a set of typical documents

Select features and options to create new profile

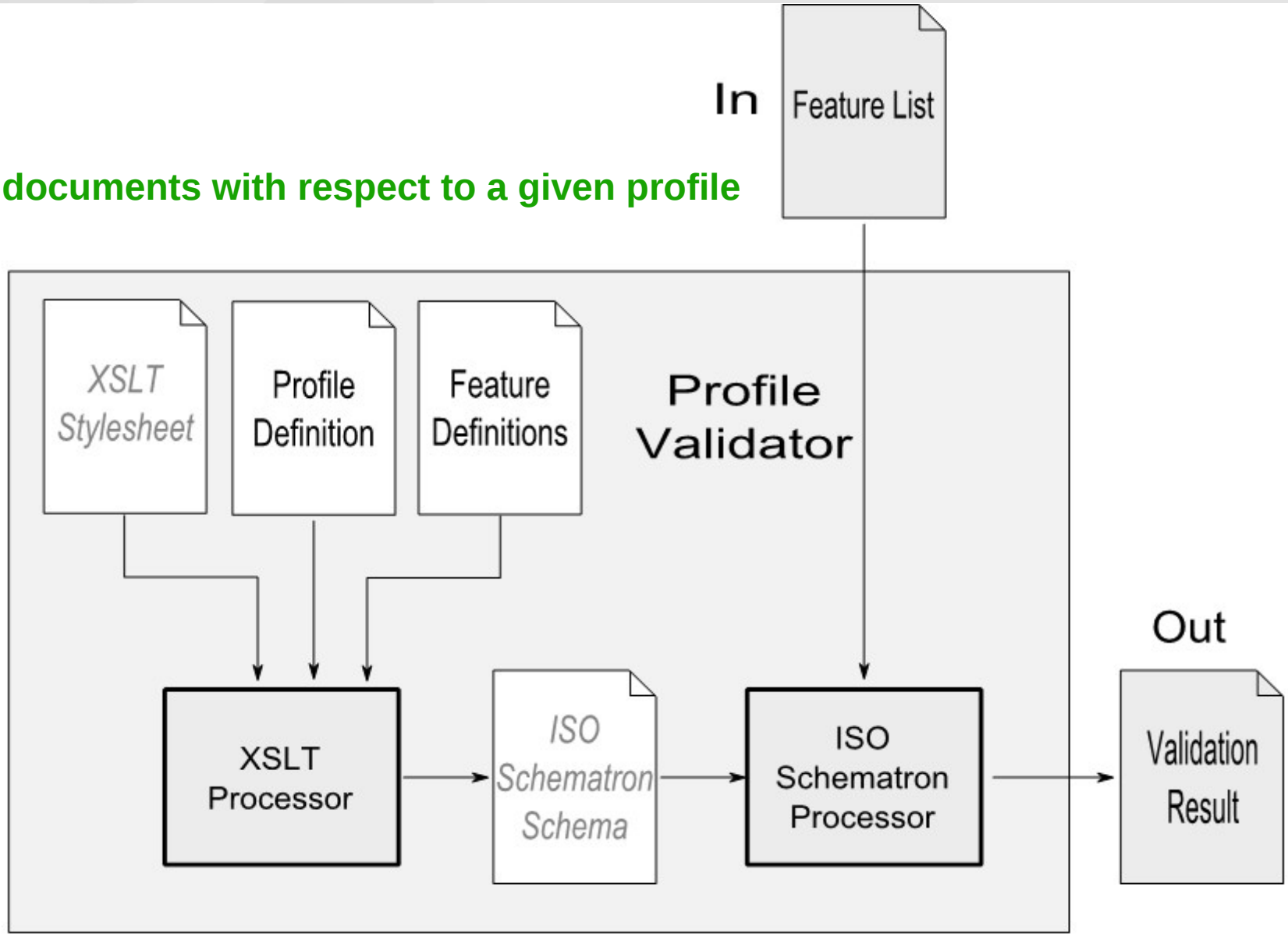
Feature Category	Feature Name	Check Feature	Relative Total Files	Total usages	Profile Feature
Absatzformatierung	Absatzformat - Ausri...	☐	0,00	2	shall
Absatzformatierung	Absatzformat - Posi...	☐	0,07	57	may
Absatzformatierung	Absatzformat - Posit...	☑	0,34	2064	should
Absatzformatierung	Absatzformat - Rah...	☑	0,33	2584	should
Absatzformatierung	Absatzformat - Links...	☐	0,92	1616	may
Absatzformatierung	Absatzformat - Scha...	☐	0,01	317	may
Absatzformatierung	Absatzformat - Zent...	☑	0,86	57040	should
Absatzformatierung	Absatzformat - Initial...	☐	0,01	19	may
Absatzformatierung	Absatzformat - Hinte...	☐	0,23	2323	may
Absatzformatierung	Absatzformat - Zeile...	☐	0,02	18	may
Absatzformatierung	Absatzformat - Zeile...	☐	0,88	124041	may

Finish editing Save profile

The Transdok Profile Validator

Fraunhofer FOKUS

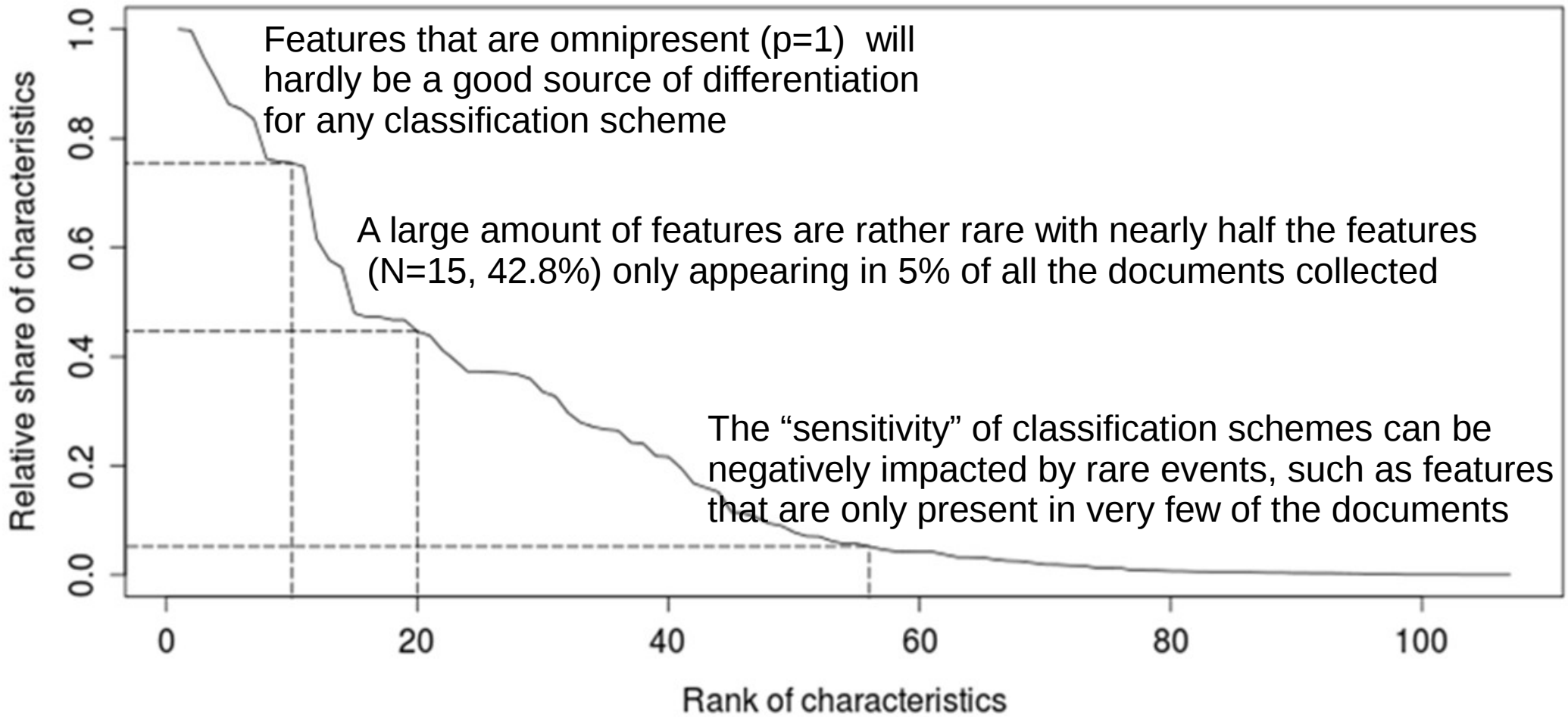
Validate documents with respect to a given profile



?! The Question !?

Fraunhofer FOKUS

Is feature based profiling a valid approach to classify documents?



Features that are omnipresent (p=1) will hardly be a good source of differentiation for any classification scheme

A large amount of features are rather rare with nearly half the features (N=15, 42.8%) only appearing in 5% of all the documents collected

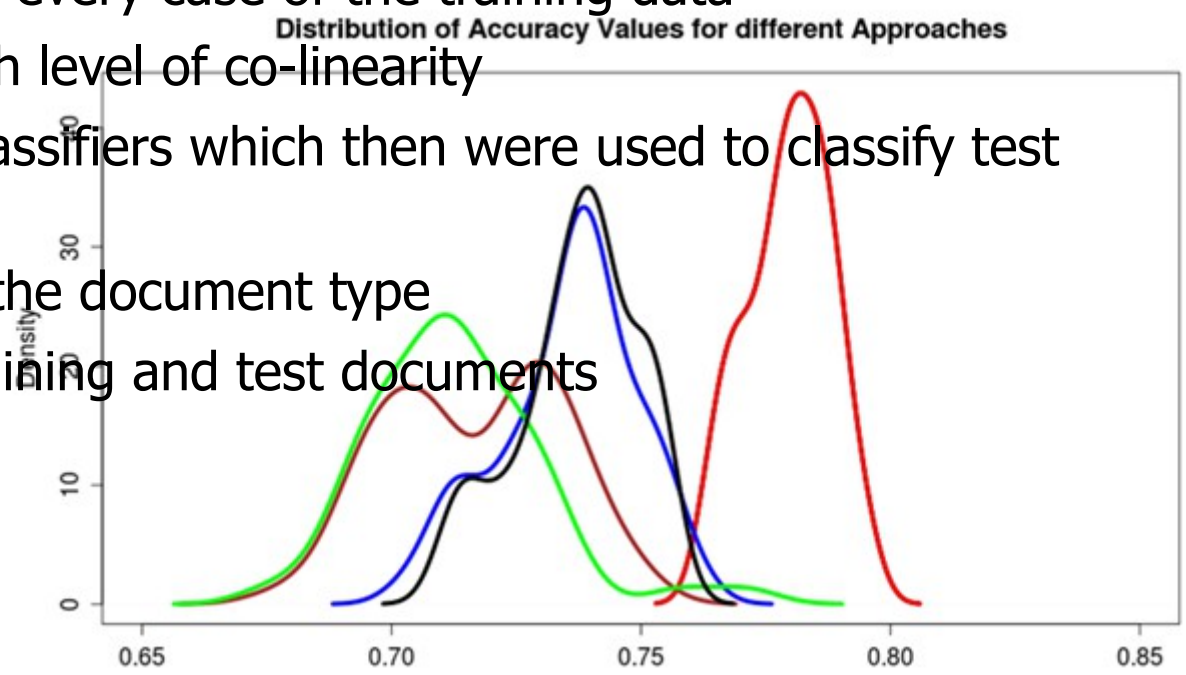
The “sensitivity” of classification schemes can be negatively impacted by rare events, such as features that are only present in very few of the documents

About Feature based Profiles

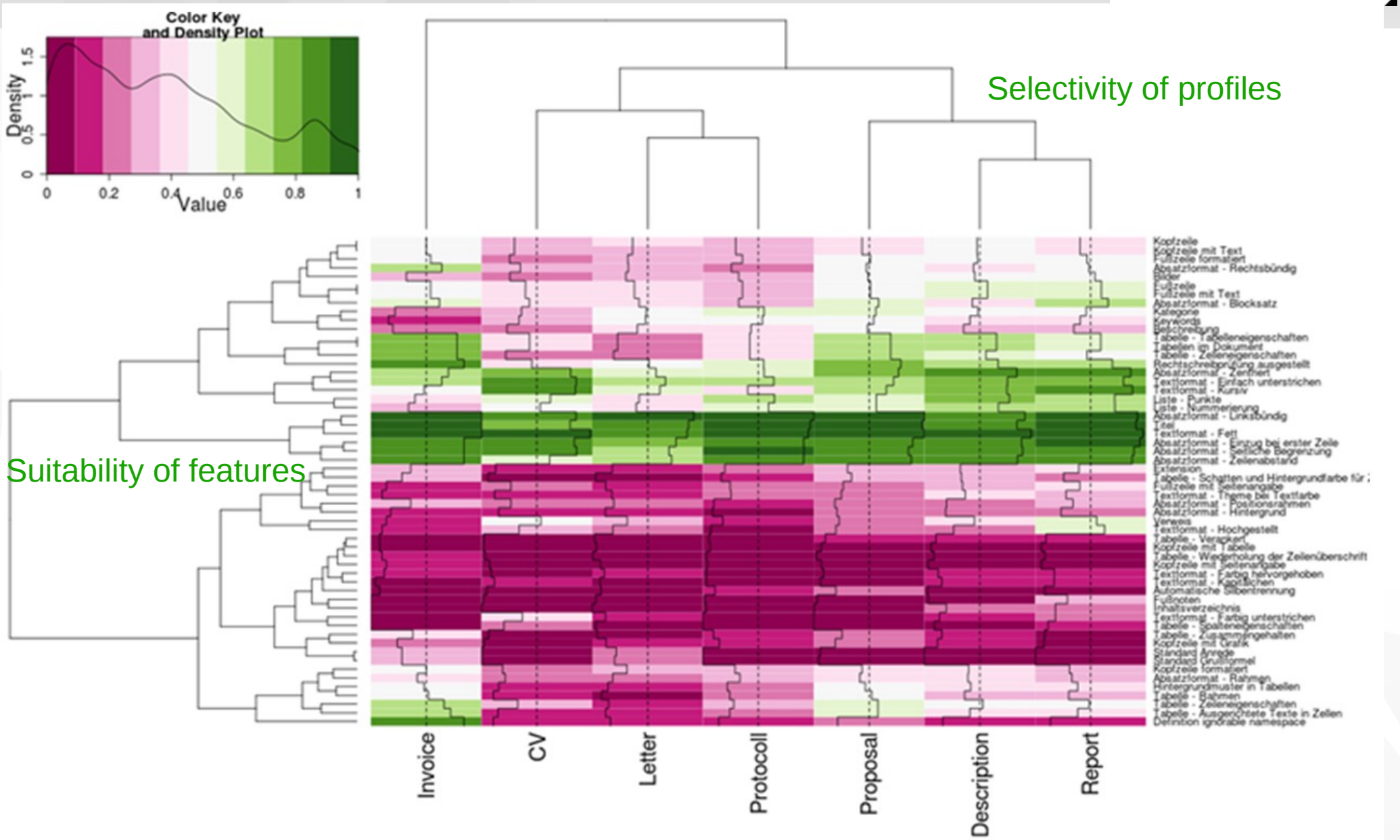
- The definition and validation of profiles depends on the suitability of the chosen features
- Desirable profile properties:
  - Profiles must be able to separate document categories
  - Documents can conform to more than one profile
  - If the definition of a profile is too weak/general, false members may be identified
  - If the definition of a profile is too strong/strict, true members may not be identified
  - The number of “false-trues” and “true-falses” should be minimal
- Binary attributes like may (may not), shall (shall not), should (should not), can (cannot) are not sufficient for the definition and validation of profiles
- Membership to a profile is not a binary decision but a matter of probability distribution

The Testbed

- Randomly select a training data set
- Shrink the number of features according to specific criteria required by a majority of classification models
  - Features that are not present in the total training data
  - Features below a threshold of occurrences in the training data
  - Features that are present for every case of the training data
  - Features which feature a high level of co-linearity
- Construct the model specific classifiers which then were used to classify test documents
- Apply the classifiers to predict the document type
- Repeat the test for different training and test documents



Cluster Analysis and Heatmaps



Conclusion

- Due to the mathematical analysis it is safe to say that some structural elements seem to exist that allow the classification of documents based on features described by XML-tags even though there is room for improvement in accuracy.
- The definition of application specific profiles for documents will improve their portability and interoperability significantly.
- It is easier to create a profile conforming document, for example using associated templates, than to check the conformity of a given document.
- The development of a reliable profile checker that is able to detect the profile having the maximum likelihood of membership seems to be a non-trivial task.

BERLIN 2012 CONFERENCE

17th-19th October

Thank you … for your attention

Any questions?

Contact:
klaus-peter.eckert@fokus.fraunhofer.de

All text and image content in this document is licensed under the Creative Commons Attribution-Share Alike 3.0 License (unless otherwise specified). "LibreOffice" and "The Document Foundation" are registered trademarks. Their respective logos and icons are subject to international copyright laws. The use of these therefore is subject to the trademark policy.