2018 13th IAPR International Workshop on Document Analysis Systems

Ontology and Framework for Semantic Labelling of Document Data and Software Methods

Christian Clausner and Apostolos Antonacopoulos
Pattern Recognition and Image Analysis Research Lab
University of Salford
United Kingdom
www.primaresearch.org

Abstract— **We present a metadata labelling framework for datasets, software tools, and workflows. An ontology for document image analysis was developed with deep support for historical data. An accompanying open source software framework was implemented to enable ontology editing, data and method annotation, workflow composition, and semantic search. A wide range of examples is used to illustrate real-world application.**

Keywords- Semantics; ontology; historical documents; document retrieval; scientific workflows; metadata

I. Introduction

With the rapidly expanding volume of data and increasing complexity of software it becomes increasingly important to establish sophisticated annotation and retrieval systems. Several metadata formats (e.g. METS [1] and Dublin Core [2]) with varying amount of detail are used to add information about authorship, publication date, technical characteristics etc. to data items such as documents and the corresponding document images.

We present a complementary framework to annotate data, but also software methods targeting specific data, using semantic labels describing the nature of the items in detail. This is part of an ongoing research project investigating scientific workflows including workflow composition systems and repositories.

Scientific workflows are used to automate software processes. To that effect, workflow systems typically offer the following features: visualisation, fault tolerance, distribution, data provenance, and repeatability. Workflows are defined by their activities (actors) and the way they are connected (data flow). Activities therein have data input and output ports.

To allow for more automation and assistive features in workflow composition and retrieval, a semantic labelling approach was developed, including a complete framework for: ontology creation and editing, workflow composition and labelling, semantic matching algorithms, and data and workflow repositories.

To the authors' knowledge, the proposed framework is the first implementation of a flexible and yet powerful semantic labelling approach (not only for documents). Other workflow systems with semantic features ([14][20]) use very strict semantic models (a workflow has to be modelled as a whole using semantic "language") and default reasoners currently limited to basic queries.

Workflow data items and online datasets have in common that a detailed semantic description helps with search and retrieval. We therefore propose that the developed ontology and labelling approach can be used to extend the existing metadata of document image datasets. The ontology includes a wide variety of concepts, complementing existing schemes which usually represent an archiving / library point of view or a very technical point of view (e.g. image attributes).

The proposed ontology and its creation are described in the next section, followed by a description of the software framework in Section III. After a discussion in Section IV the paper concludes with Section V.

II. Ontology

The ontology was developed for document image analysis and recognition, but with extensibility to other domains in mind. METHONTOLOGY [3] was used as design strategy. As the initial conceptualisation, domain-related terms were collected from various sources (text books, IJDAR papers, ACM classification scheme, project reports etc.) and put into a "term cloud". The mostly randomly arranged terms were then moved and grouped iteratively until high- and low-level concepts emerged.

To include concepts for historical material, a collection of keywords for tagging of document images was incorporated as well. They originate from the IMPACT project [4] and were refined during the Europeana Newspapers project [5]. An early description can be found in a keynote by Antonacopoulos [6]. The keywords were used to tag existing datasets [7][8] and range from document characteristics to flaws/conditions introduced through ageing, wear, or during digitisation. The original categorisation was refined to fit in with other concepts in the ontology.

The keywords are in line with related work for defects in printed documents [9][10] and photographs [11]. It is worth noting that certain aspects are modelled in more domain-specific detail in those works (e.g. grouping into physical, chemical, and biological cause of degradation or conditions specific to developed photographs), which can be used as basis for future extensions of the proposed ontology, where required.

The ontology consists of *label type* hierarchies, each targeting one aspect of a data item or an activity. A label type is thereby a concept within a class hierarchy. Labels are an instance of a label type, when attached to a data item (as metadata). In other words, the ontology is a collection of taxonomies (or partonomies), with several (more general) root label types and attached trees of label types which represent more specialised terms.

Both *activities* and *data objects* can be labelled. Here, an activity is any kind of process that transforms or

978-1-5386-3346-5/18 $31.00 © 2018 IEEE
DOI 10.1109/DAS.2018.46

generates data. A data object can be the input for or output from an activity, but also a standalone item (such as a document image) or a collection (such as a document image dataset). Partial labelling (i.e. for sub-elements of a data object) can be realised as well, but is not currently part of the implementation.

The top-level label types for activities are: domain, processing level, automation, data creation, adaptability, platform, and licence. Activities are also characterised by their input and output data, which can be labelled as well. When searching for an activity (e.g. a software tool), it can be taken into account whether it processes images of a certain kind, whether it is fully automated, whether it runs on a given operating system, and whether it produces a desired output (for instance).

The top-level types for data objects are: source, age, physical production method, acquisition method / replication steps, precision, content type, content encoding, source / target content, data granularity, data condition, data attributes, and topic. TABLE I. shows an abbreviated list of data-related label types.

The complete ontology also contains definitions and examples for terms. Currently it consists of about 350 label types. An ontology version and the option of migration rules allows for future extensions without breaking backwards compatibility. The ontology can be serialised using the OWL (Web Ontology Language) standard [12] or a simplified XML format.

The next section describes how labels are used to search for objects or aid the creation of workflows.

III. Framework

A range of algorithms and supporting software tools (Java-based) was developed to test and demonstrate the proposed labelling approach. In this section, the different components of that framework are described.

The *Ontology Editor* is used to view and edit the label types using a tree-based interface. In addition to types, label slots are used to define how many labels of one category can be assigned to a data object or activity (see Figure 1.). Export to OWL format allows the ontology to be viewed (and edited) also in other general-purpose tools such as Protégé [13] (although editing is simpler in the dedicated tool). It should be noted that the editor can be used to extend the ontology proposed in Section II or to create an entirely new one. Versioning is supported via a single integer number.

The *Workflow Editor* allows the composition of scientific workflows. The approach is loosely based on the ASKALON workflow system [14]. Different types of activities are combined to create control flow. Data flow is achieved through connecting data input and output ports of activities. The ports and also the activities can be annotated using the semantic labels.

Workflow, activity, and data repository tools (within the *Repository Hub*) are used to aggregate collections of the respective objects. They include semantic and keyword-based search functionality. The search interface is closely related to the label type hierarchies and works by gradually refining results using tick boxes (in a similar fashion to the faceted-search of online catalogues or shops).

A matching algorithm compares a given set of labels (e.g. for a data port) to a collection of label sets (e.g. from a data collection). It calculates a match score for each pair of label sets, allowing for partial matches (i.e. when a label on one side is of a higher level but similar category that a label on the other side). The more the “search labels” match the labels of a searchable object the higher is the match percentage (each non-match reducing the match score heuristically). This allows for flexible searches where the results are sorted by match score (high to low) and not restricting the search to 100% matching objects. Match details show which aspects (labels) match well and which do not. In addition to labels, also data types can be included in the search. Data type matching is strict and requires exact matches. Apart from a few primitive types (string, integer, float), data types are user-defined.

The source code and UML diagrams can be found at the authors’ GitHub profile [22].

In general, labelling and matching are used to add assistive features to the workflow framework (search, composition, validation etc.).

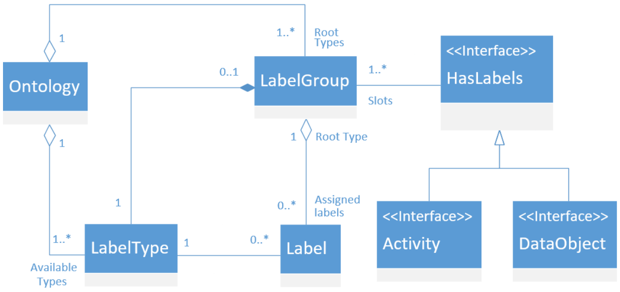

Figure 1. Ontology and labels (UML diagram). A label group represents sub-tree of the ontology for a specific root label type (i.e. a category). An object that can be labelled has slots for labels from different categories. Activities (i.e. software methods) and data objects can be labelled (have the HasLabels interface). The ontology represents the label type hierarchy but has no actual Label instances. Label instances (each having a specific label type) are attached to objects with the HasLabels interface.

IV. Discussion

This section describes general thoughts, application scenarios and labelling examples.

Although the framework allows the creation of new ontologies, better interoperability can be achieved by using the proposed ontology and extending it where necessary. Such extension activities would require central access and a consortium or a community which manages changes (including a full life cycle with versioning). Problems arising from the size of the ontology (high quantity of label types) can be solved on the user interface layer by only using a subset of labels depending on the domain of the application and by providing convenience tools such as a label quick search (e.g. by keyword). Backwards compatibility of already labelled datasets can be achieved by using migration rules that are part of the ontology itself (already implemented within the software framework).

A. Applications

Existing document image datasets could be annotated using the proposed label-based approach. Both individual items (images, ground truth files) and complete datasets can be labelled. A dataset could also have a combination of top-level labels to allow searching across different datasets, for example. Individual items could be labelled in more detail. A common use case is search and retrieval of training data for document image analysis methods. Often, data of a very specific nature is required, which, at the moment, can only be searched for by keywords (e.g. using Internet search engines).

The labelling scheme can be combined with existing description mechanisms such as METS [1], the International Image Interoperability Framework (IIIF) [15], or the PAGE format [16]. For this purpose, labels can be stored in a serialised form:

<ID Level 1>.<ID Level 2>.<ID Level 3>…

The matching and search algorithms can be applied independently from the overall software framework. Ontology versioning should be kept in mind and the version number should be stored together with the labels.

In addition to data repositories, semantic labelling can also enhance *software* and *workflow repositories* such as those of the IMPACT Centre of Competence for Digitisation in the EU [17] or the myExperiment platform [18], which typically only allow a search by keywords or by category. Semantic labels could make a significant difference in the ease of workflow composition by describing input and output data of tools as well as properties of the methods themselves.

The primary use of the proposed ontology and framework is in workflow composition and retrieval. While the developed framework is fully functional with respect to workflow composition, it does not contain a workflow execution (enactment) component. Instead of implementing such a component, the labelling approach could be ported to existing workflow systems such as Taverna [19] or Kepler [20].

B. *Examples*

This subsection provides selected examples for image conditions taken from the Europeana Newspapers Dataset [7] and the Census 1961 dataset [8]. The properties and conditions were chosen because they can have influence on recognition performance or require specialised methods.

Following selection of image characteristics and conditions from different categories are illustrated below:

- Data properties – Intended / production-related features (Figure 2.)
- Data condition – Unintended properties / flaws / issues
  - Production-related – Limitations of production method or imperfections (Figure 3.)
  - Wear/use – Problems due to (heavy) use (Figure 4.)
  - Ageing – Storage conditions or exposure (Figure 5.)
  - Acquisition- or conversion-related – Problems introduced by copying or digitisation (Figure 6.)

A given image from the Europeana set can be labelled as follows (label type hierarchy: top level – … – lower level):

- Original source – produced data – physical medium – paper document – newspaper
- Age – historical
- Physical production method – printed – typeset
- Content type – data
- Content encoding – raster image – colour
- Content of interest – visual – text
- … – visual – graphical
- Topic – Economy – financial / business
- Data granularity – physical – page
- Data properties – language – mixed languages
- … – document-related – visual – text – font – typeface class – blackletter
- … – text – font – multi-font – mixed typefaces
- … – text – font – multi-font – mixed font sizes
- … – columns – multiple
- Data condition – noisy – speckles
- … – production-related – doc. characteristics – halftoning
- … – flaws – touching characters – horizontally
- … – flaws – broken characters
- … – wear / use – medium damage – folds
- … – wear / use – additions – stamps
- … – ageing – warping
- … – ageing – discolouration
- Acquisition- / conversion-related – geometric – perspective distortions
- … – background-related – included parts / objects – paper clips
- … – method flaws – imaging-related – show-though

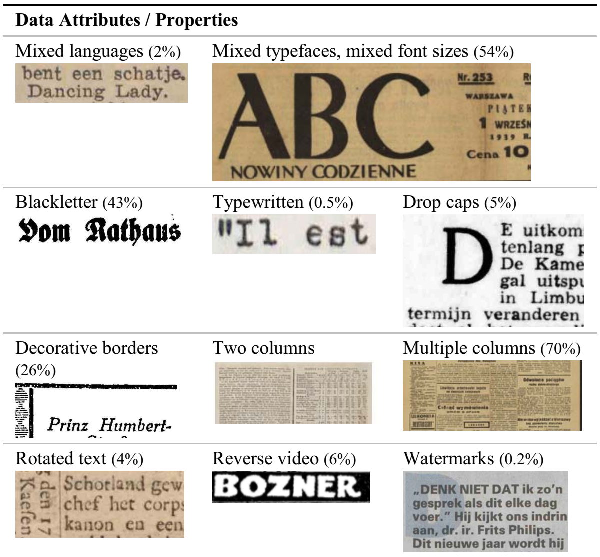
Data Attributes / Properties
Mixed languages (2%)
Mixed typefaces, mixed font sizes (54%)
Blackletter (43%)
Typewritten (0.5%)
Drop caps (5%)
Decorative borders (26%)
Two columns
Multiple columns (70%)
Rotated text (4%)
Reverse video (6%)
Watermarks (0.2%)

Figure 2. Selected image / document conditions and properties for category “Data attributes / properties”. Percentages indicate how widespread the properties are in the Europeana Newspapers set.

Finally, it should be noted that the labelling approach is not only suitable for images but also for related data such as page ground truth. A ground truth file can be labelled with:

- Original source – produced data – physical medium – paper document – newspaper
- Content type – data
- Content type – metadata

- Content encoding – structured
- Content of interest – visual – text
- Content of interest – visual – graphical
- Topic – Economy – financial / business
- Data granularity – physical – page
- Data granularity – physical – region / zone
- Data properties – language – mixed languages

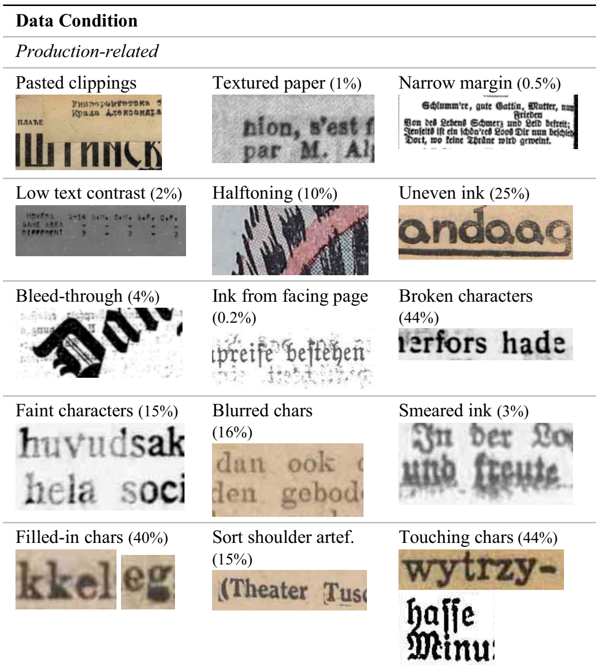

Figure 3. Selected image / document conditions and properties for category "Data condition – production-related". Percentages as in Europeana Newspapers set.

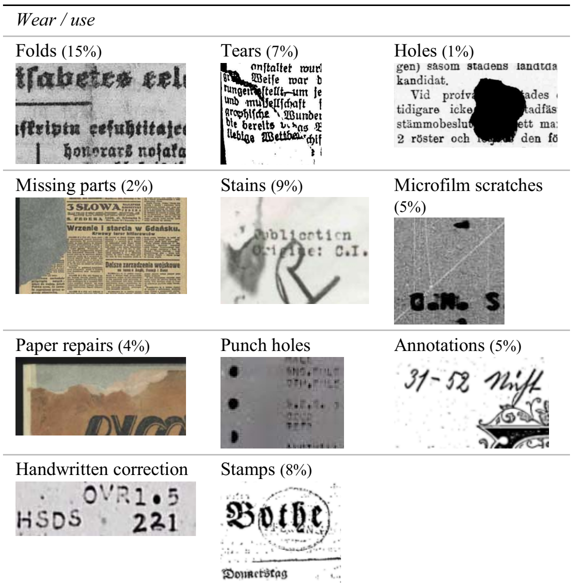

Figure 4. Selected image / document conditions and properties for category "Data condition – wear / use". Percentages as in Europeana Newspapers set.

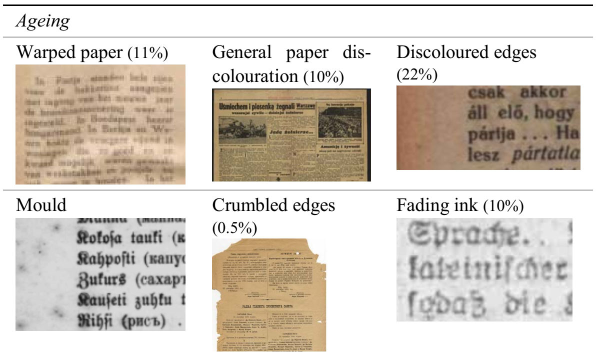

Figure 5. Selected image / document conditions and properties for category "Data condition - ageing". Percentages as in Europeana Newspapers set.

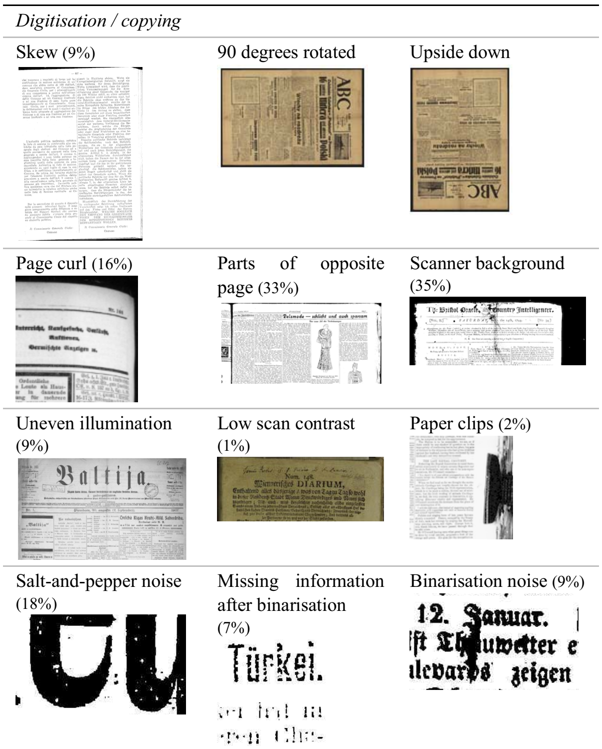

Figure 6. Selected image / document conditions for category "Data condition – digitisation / copying". Percentages as in Europeana Newspapers set.

V. Concluding Remarks and Future Work

A semantic labelling framework for data and software methods was presented. The labels describe the nature of objects in different aspects and can be used in conjunction with existing metadata formats.

The ontology, currently targeting document analysis and recognition, could be extended to other domains and complemented with further languages and scripts (e.g. using IDs from ISO standards), for instance. The intention is to offer a generic model which applies to many use cases and is open for extensions and specialisations. Further semantic concepts such as relations and literals can be added to the model if necessary.

As a demonstration, an existing public dataset (the Europeana Newspapers set [7], for instance) might be labelled using the proposed scheme and an online interface can be developed to access the information.

The complete version of the presented ontology and the software framework are available on GitHub [22].

TABLE I. DATA OBJECT LABELS (235 OUT OF 350 TOTAL LABELS FOR DATA AND SOFTWARE METHODS). HIGH-LEVEL TYPES ON THE LEFT, LOWER-LEVEL TYPES ON THE RIGHT.

Lev.1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	Lev. 6	Lev. 7
Original Source						
	Produced data					
		Physical source medium				
			Paper document			
				Book		
				Newspaper		
	Captured data					
		Real / natural scenes				
Age						
	Historical					
		Medieval				
	Contemporary					
	Ancient					
Physical Production Method						
	Manual					
	Machine					
		Printed				
			Typeset			
			Computer printout			
		Typewritten				
Acquisition / Replication Method						
	Analog / physical to digital					
		Scanning				
		Camera				
	Copied					
		Photocopy				
		Microfilm / microfiche				
	Synthesis					
Precision						
	Ground Truth / gold standard					
	Measured					
	Estimated					
	Random					
	Fuzzy					
Content Type						
	Data					
	Metadata					
		Quality				
			Performance Information			
		Features				
		Structure				
			Table of contents			
		Annotations				
		Authorship				
		Spatial				
			Location			
	Settings					
	Model					
	Lexicon / index					
	Corpus / database					
Content Encoding						
	Textual					
		Annotated				
	Structured					
	Raster image					
		Colour Image				
		Bitonal				
	Mathematical / geometrical					
		Vector-based				
			Stroke-based			
		Polygonal				
Content of Interest						
	Visual content					
		Text				
		Graphical				
			Separators			
			Barcode / QR Code			
		Image				
			Photograph			
				Person(s)		
			Drawing			
		Mixed / composite content				
			Tables / forms			
			Charts			
			Maps / plans			
			Mathematical expression			
Data Granularity						
	Physical / visual granularity					
		Document-related				
			Double-page			
			Page			
			Region / Zone			
			Text line			
			…			
		Natural language-related				
			Sentence			
			Token / chunk			
			Syllable			
	Logical granularity					
		Document-related				
			Document			
			Chapter			
			Section			
			Article			
			Paragraph			
Data Condition						
	Noise					
		Speckles				
			Salt-and-pepper noise			
		Clutter				
			Thresholding-related noise			
	Production-related					
		Document characteristics				
			Pasted clippings			
			Textured paper			
			Uneven character spacing			
			Narrow border			
			Low paper-to-content contrast			
			Halftoning			
			Dithering			
		Document faults				
			Bleed-through			
			Ink from facing page			
			Smeared ink			
			Touching characters			
				Horizontally		
				Vertically		
			Uneven ink distribution			
			Filled-in characters			
			Sort shoulder artefacts			
			Broken characters			
			Faint characters			
			Blurred characters			
			Non-straight text lines			
	Wear / use					
		Medium damage				
			Folds			
			Tears			
			Holes			
				Punch holes		
				Unintended holes		
			Missing parts			
			Stains			
			Scratches			
			Staples			
		Additions				
			Visible repairs			
				Paper repairs		
				Clear tape		
			Informative			
				Annotations		
				Stamps		
			Corrections			
				Manual corrections		
	Ageing					
		Warping				
		Discolouration				
			Global			
			Edges			
		Disintegration				
			Uneven edges			
		Mould				
		Fading content				
	Acquisition / conversion-related issues					
		Geometric issues				
			Skew			
				Global		
				Non-uniform		
			90-degree rotation			

			Upside down			
			Perspective distortions			
			Page curl			
		Content / background-related				
			Incomplete capture			
			Tight / narrow margins			
			Included other objects			
				Part of pre- or succeeding object		
				Medium structure (book cover…)		
				Paper clips		
				Fingers		
				Insects		
				Background (e.g. scan bed)		
		Method flaws				
			Imaging-related			
				Show through		
				Uneven illumination		
					Shadows	
				Out-of-focus		
				Low contrast		
				Missing / changed content		
					Due to thresholding	
Data Attributes / Properties						
	Language					
		Natural language				
			English			
		Mixed languages				
	Document-related					
		Visual properties				
			Text-related			
				Script		
					Braille	
					Latin	
				Font-related		
					Cursive	
					Monospace	
					Hand / Typeface class	
						Blackletter
						Antiqua
						Medieval manuscript
					Decorated text	
						Flourishes
						Multiple colours
						Reverse video
					Multi-font	
						Mixed typefaces
						Mixed font sizes
				Drop caps		
			Columns			
				One column		
				Two columns		
				Multiple columns		
			Rotated content			
			Complex background			
				Watermarks		
				Impressions / embossing		
			Illustrations			
				Multi-coloured		
			Decorations			
				Frames / borders		
			Line drawing / line-art			
			CAPTCHAs			
		Structural				
			Running titles			
			Footnotes			
			Bibliographic reference			
Topic						
	Economy					
		Financial / business				
			Bank checks			
			Invoices			
	Social science / environmental					
		Maps				
			Topographical maps			
			Road maps			
		Traffic and automotive				
			Number plates			
			Traffic signs			
	Science and Engineering					
		Architectural				
			Floor plans			
			Architectural drawings			
		Medical				
		Engineering drawings				
		Patents				
	Media / entertainment					
		Advertisements				
	Computing					

References

[1] J. P. McDonough, “METS: standardized encoding for digital library objects” Int. Journal on Digital Libraries 6, 2, pp. 148–158, Apr. 2006, DOI: https://doi.org/10.1007/s00799-005-0132-1

[2] Dublin Core Metadata Initiative, http://dublincore.org/, accessed 09/08/2017

[3] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, “METHONTOLOGY: From Ontological Art Towards Ontological Engineering”, AAAI-97 Spring Symposium Series, 1997, American Association for Art. Int.: Stanford University, EEUU.

[4] IMPACT Project 2011 Publishable summary. http://www.impact-project.eu/documents , 2012.

[5] Europeana Newspapers Project., Staatsbibliothek zu Berlin, http://www.europeana-newspapers.eu, [cited 2016 01/06]

[6] A. Antonacopoulos, “Large-Scale Digitisation and Recognition of Opportunities for Image Processing and Analysis Historical Documents: Challenges and Analysis”, Keynote presentation given at the Swedish Symposium on Image Analysis 2010 (SSBA2010), Uppsala, Sweden, March 11-12, 2010.

[7] C. Clausner, C. Papadopoulos, S. Pletschacher, A. Antonacopoulos, “The ENP Image and Ground Truth Dataset of Historical Newspapers”, Proc. 13th Int. Conf. on Document Analysis and Recognition (ICDAR2015), Nancy, France, 08/2015, pp. 931-935.

[8] C. Clausner, J. Hayes, A. Antonacopoulos, S. Pletschacher, “Unearthing the Recent Past: Digitising and Understanding Statistical Information from Census Tables”, Proc. Second International Conference on Digital Access to Textual Cultural Heritage (DATeCH), Goettingen, Germany, 01-02 June 2017.

[9] R. D. Lins, “A Taxonomy for Noise in Images of Paper Documents - The Physical Noises”, Proceedings of the 6th Int. Conf. on Image Analysis and Recognition, Halifax, Canada, July 6-8, 2009.

[10] H. S. Baird, “The State of the Art of Document Image Degradation Modelling”, Book chapter in Digital Document Processing: Major Directions and Recent Advances. Springer, Lon., 2007, p. 261-279.

[11] E. Ardizzone, A. De Polo, H. Dindo, G. Mazzola, C. Nanni, “A Dual Taxonomy for Defects in Digitized Historical Photos”, 10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009.

[12] W3C, OWL 2 Web Ontology Language Primer (Sec. Ed.). 2012

[13] Protégé - A free, open-source ontology editor and framework for building intelligent systems. [cited 2015 24/03].

[14] J. Qin, T. Fahringer, “Scientific workflows : programming, optimization, and synthesis with ASKALON and AWDL”, 2012, Berlin, New York: Springer. xxi, 222 pages.

[15] International Image Interoperability Framework – IIIF, http://iiif.io

[16] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, 2010, pp. 257-260.

[17] IMPACT Centre of Competence for Digitisation, https://www.digitisation.eu

[18] D. Roure, C. Goble, R. Stevens, “The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows”, Future Generation Computer Systems, 2008. 25: p. 7.

[19] T. Oinn et al., “Taverna: a tool for the composition and enactment of bioinformatics workflows”, Bioinformatics, 2004. 20(17): p. 3045-3054.

[20] B. Ludaescher et al., “Scientific workflow management and the Kepler system: Research Articles”, Concurr. Comput. : Pract. Exper., 2006. 18(10): p. 1039-1065.

[21] Y. Gil et al., “Expressive reusable workflow templates”, e-Science, 2009, e-Science'09. Fifth IEEE International Conference on. 2009. IEEE.

[22] Semantic labelling software framework GitHub repository, https://github.com/PRImA-Research-Lab/semantic-labelling