Glossary for the Mass Digitisation of Text & OCR

IMPACT Workflow Resource

IMPACT project
Niall Anderson, British Library

Table of Contents

Introduction

The IMPACT Glossary for the Mass Digitisation of Text & OCR covers many terms in general use that relate to all stages in a workflow of text-based digitisation. However the usage of some general terms may be influenced by IMPACT's work. For instance, the definition of the term "Optimisation" presumes that any optimising of images will be done with the aim of improving OCR. Where definitions have been arrived at by reference to external sources, original spellings have been retained.

Knowledge Base

Term	Definition	Reference
Access Image	**A low resolution image made available to service users for reference or use. An access image is usually derived from a higher resolution master image.**	
Adaptive OCR	**Adaptive Optical Character Recognition automatically adapts itself to the qualities and format of the text being recognised. Adaptivity can be focused on text content and/or font shape. It may be driven either by manual data entry results (see: Ground Truth) or by automatic computation.**	

ALTO	**Abbreviation for Analyzed Layout and Text Object.**	
Analyzed Layout and Text Object	**ALTO is an extension to the Metadata Encoding and Transmission Standard (METS). Created as part of the METAe Project and commercialised by project partner Content Conversion Specialist, GmbH. ALTO is now maintained by the Library of Congress Network Development and MARC Standards Office. ALTO describes the layout and content of text-based digital items in detail, including references to character coordinates within a digital space. See also: Metadata Encoding and Transmission Standard (METS).**	ALTO Schema, Library of Congress [http://www.loc.gov/standards/alto/alto-v2.0.xsd]
Authority Control	**The consistent use and maintenance of the forms of names, subjects, uniform titles, etc., used as headings in a catalogue. The process creates a link between the bibliographic record of an individual item in the library and the terms used to describe its content. See also: Named Entities.**	British Library [http://www.bl.uk/bibliographic/authority.html]
Authority File	**An index of decisions made about how names, concepts, serial publications, etc., are referred to in a library catalogue or other knowledge repository. By using a controlled vocabulary, knowledge repositories can group together similar items and also explain their similarity, allowing for greater usability. See also: Named Entities**	British Library [http://www.bl.uk/bibliographic/authority.html]
Batch Processing	**A system whereby digital files are processed en masse rather than individually, thus requiring less operator oversight.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary/#b]

Batch sample	**In mass digitisation, batch sampling is a means of quality assurance whereby a numerically significant subset of a larger body of digital information is taken as being qualitatively representative of the whole. Where the subset is taken as being of acceptable standard (in terms of text legibility, OCR accuracy, etc.), the entire batch will be passed; where the subset is taken as being unacceptable, the entire batch will be rejected.**	
Benchmarking	**A sampling of representative files or tasks within a workflow, used to set quality targets and to measure subsequent work against.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary/#b]
Binarisation	**Binarisation is a process that converts a colour or grey scale image into a black and white image. Image binarisation is applied before OCR, because the contrast between black and white allows an OCR engine to more easily distinguish significant detail from background. See also: Binary, Binary Image**	
Binary	**Anything which is binary has only two possible values. All computer data is represented in binary form, as a series of bits which have the value 0 or 1. See also: Binarisation; Binary image.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary/#b]
Binary image	**A binary image is either: 1. an image stored in binary format (i.e. as a series of 0/1 values) 2. an image with only two distinct tones or colours. See also: Binarisation; Binary.**	
Bit	**An abbreviation for 'binary digit': the smallest unit of digital information within a computer system, represented as a 1 or 0. See also: Binary.**	Langford's Advanced Photography, Seventh Edition (2007), Focal Press, p. 303

Bit Depth	**The bit depth of an image refers to the number of bits used to describe the colour of each pixel. Greater bit depth allows more colours to be used in the colour palette for the image. 1-bit per pixel will allow black and white, 8-bits per pixel will allow 256 colours, 8-bits per colour component in an RGB image (24 bit) will allow 16777216 colours.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary/#b]
Bitmap	**An image formed by a grid of pixels. A computer assigns a value to each pixel - ranging from 1 'bit' of information (simply black or white) to as many as 24 bits per pixel for full colour images. Also sometimes known as a pixmap.**	Langford's Advanced Photography, Seventh Edition (2007), Focal Press, p. 303
Bitonal image	**Bitonal denotes the existence of two distinct tones/colours within images. Typically the two tones/colours used for a bitonal image are black and what, though in principle it could refer to any two colours.**	
Black and white image	**A black and white image is a special case of a Bitonal image containing the two colours black and white.**	
Bleed Through	**An artefact of printing when paper is lightweight or the ink heavily applied the colour can bleed through to the other side. This can have a severe negative effect on OCR accuracy.**	
Boutique digitisation	**The planned and ad hoc conversion of small amounts of analogue material, often unique and of high value, into high-quality digital format. Selection of volumes is usually precise, manual input is high, and the workflow is flexible so it can be adapted to new conditions. Work in this area is most often project based.**	
Calibration	**The process of adjusting the colour of one device relative to another, such as a monitor**	

	to a printer, or the process of adjusting the colour of one device to an established standard. See also: OCR accuracy; Colour management.	
Capture	**The process of obtaining a digital image from a vision sensor, such as a camera or scanner.**	John Daintith, A Dictionary of Computing (2004), Oxford University Press.
CENL	**Abbreviation for Conference of European National Libraries.**	Foundation Conference of European National Libraries [http://www.nlib.ee/cenl/index.php]
Characterisation	**In imaging, characterisation is the difference between the actual representation of colour within a digital image and the intended representation.**	
CMYK	**Cyan, Magenta, Yellow and Black. The first three are the basic subtractive colour dyes formed in most colour emulsions and printers' inks. Pure black is unobtainable by combining CMY inks, so is generally added as a fourth colour in printing to improve image body and contrast. See also: RGB.**	Langford's Advanced Photography, Seventh Edition (2007), Focal Press, p. 303
Collaborative correction	**Collaborative correction refers to the post-publication correction of OCR results by a community of users.**	
Colour Depth	**See: Bit depth.**	
Colour Management	**Procedures and/or electronic systems designed to ensure uniformity of colour appearance within digital images across input, monitor and output devices. See also: Calibration**	Langford's Advanced Photography, Seventh Edition (2007), Focal Press, p. 304
Colour model	**An abstract system for representing colours as ordered sets of numbers in a limited set. An example is the RGB colour model, where colours are assigned numbers based on the degree to which they are interpreted as red, green or blue, and in what combination. See also: Colour management**	

Colour profile	**A technical statement of how an input, monitoring or output device will interpret and present colour in a digital image. Such devices are often naturally tuned to slightly different colour specifications (known as colour spaces), so colour management procedures and systems must be used to ensure uniformity of colour across all devices. See also: Colour management; Colour space.**	
Colour space	**A representation of all possible mixtures of the primary colours within a given colour model, with each variant assigned a unique place within the representation. See also: Colour management; Colour model; Colour profile.**	
Colour target	**A reference chart of colour patches with known properties, which can be used to calibrate image capture devices so that the colour displayed to the end user matches the colour of the original. See also: Compression.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/stillimages/advice/colour-and-resolution-targets/]
Compression	**Compression of digital image data to reduce storage requirements or transmission speed across networks. See also: Lossless compression; Lossy compression.**	
Computer vision	**Computer vision is the theory underpinning the design of artificial systems that can obtain information from images. Optical Character Recognition is a discipline within computer vision. See also: Optical Character Recognition.**	
Confidence Level (OCR)	**OCR confidence is a numerical value assigned by an OCR engine to a text component, indicating the degree to which the engine is certain that it has recognised the component correctly. The confidence**	

	level can apply to any text component, from single characters to complete pages or documents. See also: Optical Character Recognition	
Continuous Tone image	**An image with continuous tone is one who pigmentation - or ink - varies in concentration to the degree that the tones in the image are strong or weak. Saturated tones mean a high concentration and light tones mean a low concentration. The most common form of continuous tone images are photographs. See also: Half Tone image; Bitonal image.**	
Dataset	**An accumulation of data products, secondary or ancillary data and software, and documentation that record and support the use of those data products. Within IMPACT, the term dataset refers to an image collection, with associated metadata and ground truth. See also: Ground truth; Metadata**	NASA [http://pds1.jpl.nasa.gov/documents/dpw/appc.html]
Density	**The range of tones that can be captured by an imaging device. Optical density is measured on a scale of 0 for white to 4 for black.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary#optical-density]
Derivative image	**An image created from a digital master image, usually compressed to preserve storage space and shorten transmission time across a network. See also: Access image; Compression.**	
Deskewing	**Correction of distortion/ rotation caused by image capture from a viewpoint other than on the perpendicular.**	JISC Digital Media [http://www.jiscdigitalmedia.ac.uk/glossary#optical-density]
Dewarping	**The post-capture manipulation of a digital image that corrects for the warped surface of an original page. Dewarping changes the orientation of text within a digital image until it appears as straight-line text,**	

	thus making it readable to OCR engines.	
Digital Record Object Identification (DROID)	**DROID is a software tool developed by the UK National Archives that performs automated batch identification of file formats. It is designed to meet the fundamental requirement of any digital repository to be able to identify the precise format of all stored digital objects and link that identification to a central registry of technical information about that format and its dependencies. It uses internal and external signatures to identify and report the specific file format versions of digital files. These signatures are stored in a XML signature file generated from information recorded in the PRONOM technical registry. New and updated signatures are regularly added to PRONOM, and DROID can be configured to automatically download updated signature files from the PRONOM website via web services. See also: PRONOM.**	DROID website [http://droid.sourceforge.net/]
Document Type Definition (DTD)	**An XML language that describes a document using a set of declarations/definitions that conform to a particular markup system. See also: Markup language; Schema.**	DTD at W3C [http://www.w3.org/TR/REC-xml/#dt-doctype]
Dots per inch (DPI)	**DPI is an artefact of spatial printing whereby what appears to the viewer as a continuous image is in fact made up of many tiny ink-dots close together. The number of dots per inch tends to correlate with the resolution of a digital image (with high DPI implying high resolution), but because DPI is a characteristic of printing rather than capture, the relation is only indirect. See also: PPI.**	Wikipedia [http://en.wikipedia.org/wiki/Dots_per_inch], DPI vs. PPI [http://www.tildefrugal.net/photo/dpi.php]

DTD	**Abbreviation for Document Type Definition.**	
Dublin Core (DC)	**Dublin Core is a metadata element set created to provide a fundamental body of elements that can be shared across disciplines or within any type of organisation needing to organise and classify information. The "core" consists of a set of property types and values that provide semantic information about the nature and content of web resources, much the same way a library card catalogue provides indexes of book properties. In technical terms, Dublin Core is a simple standardised metadata element set, using XML and RDF for cross-domain information resource description. See also: Metadata**	Dublin Core Metadata Initiative [http://dublincore.org/]
Emulation	**In digital preservation, emulation is a technical strategy employed to artificially replicate the environment in which a digital object was originally created. Where digital objects have significant executable or interactive elements (a computer game, for instance; or a legacy metadata wrapper), it may be safer and more cost effective to recreate an object's original technical platform than to migrate individual object to a new format. See also: Migration.**	NOF Digitisation Guidelines [http://www.ukoln.ac.uk/nof/support/manual/digital-preservation/intro.htm#Technology%20emulation]
Extensible Markup Language (XML)	**The Extensible Markup Language (XML) is a general-purpose specification for creating custom markup languages. It is classified as an extensible language, because it allows its users to define their own elements. Its primary purpose is to help information systems share structured data, particularly via the Internet, and it is used**	XML at W3C [http://www.w3.org/XML/], XML 1.1 Recommendation [http://www.w3.org/TR/xml11/]

	both to encode documents and to serialize data. See also: Markup language; Metadata	
Font	**In typography, a font is defined as a single complete character set in a particular size and style. For instance, Times New Roman Italic 12 is a different font from Times New Roman Bold 11, despite both sets sharing the same typeface. See also: Typeface.**	
Gamma	**A numerical system used to quantify the amount of contrast in an image. Gamma correction is the process by which the contrast of an original image can be altered to suit the specifications of a particular output device or standard.**	
Granularity	**The granularity of an image is measured as the smallest amount of magnification necessary to produce a grainy effect in the image. Granularity is related to the relative density of the inks in a printed image, with low granularity signalling a high density of ink.**	Stroebel, Leslie, View Camera Techniques
Grayscale (Gray Scale)	**See under: Greyscale.**	
Greyscale (Grey scale)	**A number of greys ranging from black to white. An eight bit greyscale image could have 254 greys between black and white. Greyscale images are distinct from black-and-white images, which in the context of computer imaging are images with only two colours, black and white. See also: Bitonal.**	
Ground Truth	**In digital imaging and OCR, ground truth is the objective verification of the particular properties of a digital image, used to test the accuracy of automated image analysis processes. The ground truth of an image's text content, for instance, is the complete and accurate record of every character and word in the**	

	image. This can be compared to the output of an OCR engine and used to assess the engine's accuracy, and how important any deviation from ground truth is in that instance.	
Half Tone image	**A half tone image is one where an apparently continuous image is in fact comprised of a concentration of single dots of different tones and/or colours. A common form of half-tone image is a dot matrix printout. See also: Continuous Tone image, Bitonal image**	
HSV	**An abbreviation for Hue, Saturation and Value. HSV is a standard colour model that describes colour in terms tint, narrowness of colour spectrum and intensity. See also: Colour model.**	
Image collection	**Within IMPACT, the term image collection refers to a comprehensive collection of images with their associated metadata and ground truth. These image collections have been created as a means of testing and evaluating the performance of techniques and tools used during the text recognition process.**	
Image compression	**Compression refers to any algorithm applied to a digital image to reduce its file size, enabling swifter transfer across a network the preservation of storage space. Compression techniques are distinguished by whether they remove detail and colour from the image. Lossless techniques compress image data without removing detail; lossy techniques compress images by removing detail. See also: Lossless compression; Lossy compression.**	Australian Museums and Galleries Online [http://archive.amol.org.au/capture/course/glossary.html]
Image Enhancement	**The post-capture improvement of the quality of scanned documents. Within IMPACT,**	

	image enhancement is carried out by a suite of software designed to maximise OCR results by correcting for underlying faults in the image and by optimising the presentation of the image for machine readability.	
Industrial digitisation	**A digitisation workflow of any size that operationalises and integrates digitisation processes. It is characterised by the automated scanning, description, processing and publication of digital resources. Operationalisation of these steps allows the institution to derive increased strategic benefits which may be more difficult to achieve in a project based environment.**	
Information retrieval	**Information retrieval is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web.**	Wikipedia [http://en.wikipedia.org/wiki/Information_retrieval]
IPR	**Abbreviation for Intellectual Property Rights. Intellectual property covers an individual or company's moral right to ownership and the right to earn money from the product of intellectual activity in a commercialised field.**	WIPO Intellectual Property Handbook [http://www.wipo.int/about-ip/en/iprm/index.html]
JHOVE	**Abbreviation for the JSTOR/ Harvard Object Validation Environment, a joint project of JSTOR and the Harvard University Library to develop an open source framework for format identification, validation and characterisation.**	JHOVE [http://hul.harvard.edu/jhove/]
Language model	**Language modelling is a practice within artificial intelligence whereby a computer is trained to recognise - and to a limited extent interpret - words by a probabilistic analysis of a**	

	large body of text in a given language. Once the corpus has been fed into a computer, a bespoke script will run that determines the overall number of instances of a particular word and also the words that it is most/least often associated with in a phrase. Language modelling can thus be used to refine the OCR process, by excluding those words that are statistically least likely to be associated with one another.	
Large-scale digitisation	**The conversion of large amounts of content into digital format. Like mass digitisation, large-scale digitisation is characterised by structured and largely automated workflows. Because of the smaller amount of material being processed, large-scale digitisation may include more flexible criteria for selection, more manual input, and some value-added features such as the structured tagging of OCR output.**	
Lexicon building	**In linguistics, the lexicon of a language is its vocabulary, expressed as units - "lexemes" - that correspond to the forms of particular words. These forms may be linked semantically (as in the relationship of meaning between the words "have", "has" and "had"), or grammatically (that is, as belonging to a particular word category or part of speech), or historically (as in cases of spelling variation across time or between countries). Within IMPACT, lexica have been built for several languages to both aid general linguistic research and to support the OCR process by increasing the number of words in an OCR engine's dictionary.**	
Lossless compression	**A compression algorithm that reduces the storage space**	Australian Museums and Galleries Online [http://

	needed to house an image file, without real loss of data. The uncompressed image can be reconstructed to be identical to the original. Continuous tone images will on average be reduced to half the original size. See also: Compression; Lossy compression.	archive.amol.org.au/capture/course/glossary.html]
Lossy compression	**A compression algorithm that reduces file size by actually removing data from the image. The most effective lossy algorithms work by discarding information that is not easily perceptible to the human eye. Effective compression ratios of 10:1 to 50:1 can be attained. See also: Compression; Lossless compression.**	Australian Museums and Galleries Online [http://archive.amol.org.au/capture/course/glossary.html]
MARC	**Abbreviation for Machine-Readable Cataloging - a standard bibliographic format used in libraries.**	JISC Digital Media
Markup language	**A markup language is a system for annotating a text or other knowledge resource in a manner that is syntactically distinguishable from the text or main content of the resource. Electronic markup languages generally distinguish content from content type (title, author, data, etc.) by presenting the content type as standardised bracketed tabs. See also: Metadata; Extensible Markup Language (xml)**	
Mass digitisation	**The conversion of extremely large amounts of printed material into digital format. Volumes in a mass digitisation workflow are selected on the basis of shared characteristics (bibliographic, rights-related, physical, etc.) to optimise the speed with which they can be scanned, described and published. Mass digitisation is characterised by a high degree of automation, highly**	

	structured workflows and low manual input. OCR tends to be used as an index to the basic metadata. The most prominent example of a mass digitisation project is Google Books.	
Master image	**A digital image, also referred to as a master image, that has been captured at the highest practicable quality or resolution, usually for long-term usage. Archival images of this sort are normally stored in an offline mode and are accessed only for the production of surrogate or derivative images.**	Australian Museums and Galleries Online [http://archive.amol.org.au/capture/course/glossary.html]
Metadata	**Data about the content, format and use of a knowledge resource, compiled to enhance the discoverability of the resource in a digital space, and to give information about the technical and legal standards that govern its use. Metadata therefore usually includes information about the intellectual content of a resource, digital representation data, and security or rights-management information.**	Australian Museums and Galleries Online [http://archive.amol.org.au/capture/course/glossary.html]http://en.wikipedia.org/wiki/Metadata
Metadata Encoding & Transmission Standard (METS)	**The Metadata Encoding & Transmission Standard is an XML-standard maintained by the Library of Congress for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.**	Library of Congress: METShttp://www.loc.gov/standards/mets/ [http://www.loc.gov/standards/mets/]
Metadata for Images in XML Schema (MIX)	**Metadata for Images in XML Schema is an XML schema maintained by the Library of Congress for a set of technical data elements required to manage digital image collections.**	Library of Congress: MIX [http://www.loc.gov/standards/mix/]
Metadata Object Description Schema (MODS)	**The Metadata Object Description Schema is a schema maintained by the Library of Congress for a bibliographic element set that may be used for a variety of**	Library of Congress: MODS [http://www.loc.gov/standards/mods/]

	purposes, and particularly for library applications.	
Metrics	**Metrics are a system of parameters or ways of quantitative and periodic assessment of a process that is to be measured.**	Wikipedia [http://en.wikipedia.org/wiki/Metrics]
METS	**Abbreviation for Metadata Encoding & Transmission Standard.**	
Microfilm	**First generation microfilm is produced by microfilming from original, usually print, sources. First generation films are used for long term preservation and to produce second generation films. Second generation microfilm is produced by making a copy from a first generation microfilm before it is stored for preservation. Second generation films are used to produce user copies ('third generation' films) on microfilm or microfiche, and more recently for making digital copies through scanning. Second generation films can have either a positive or negative polarity. See also: Microform polarity.**	
Microform polarity	**A microform is any photographic storage medium that requires magnification to be read by the human eye. Examples include microfilm, microfiche and photographic film. Microforms can have positive or negative polarity. In positive polarity microforms, the colours and tones of the image will be represented as they appeared in the original object. In negative polarity microforms, the colours and tones are reversed, with light areas appearing dark and vice versa. Negative images have much less natural contrast than positive images, but the contrast of negative images**	

	can be greatly enhanced when the microform is scanned and digitally post-processed - often producing cleaner and more easily machine readable images.	
Migration	**The transfer of digital information across software and hardware generations to preserve data integrity and future access. See also: Emulation.**	
MIX	**Abbreviation for Metadata for Images in XML Schema.**	
MLA	**Corresponds to the community of Museums, Libraries and Archives.**	The Museums, Libraries and Archives Council [http://www.mla.gov.uk/home]
MODS	**Abbreviation for Metadata Object Description Schema.**	
Named Entities (NE)	**Named entities are text units such as the names of persons, organisations, locations, etc. Named entities present a particular problem to OCR accuracy, in that the internal dictionaries of OCR engines tend to recognise only the most common names in each category. See also: Named Entity Recognition; OCR accuracy; Authority files.**	Wikipedia [http://en.wikipedia.org/wiki/Named_entity_recognition]
Named Entity Recognition (NER)	**Named Entity Recognition is a task within information extraction that seeks to locate and classify named elements in text. Within IMPACT, NER is used to enhance the accuracy of OCR results by expanding the dictionary of the IMPACT Adaptive OCR system to include the names of people, organisations, locations, etc., in a variety of languages. See also: Named Entities; OCR accuracy; Authority files.**	
NE	**Abbreviation for Named Entities.**	
NER	**Abbreviation for Named Entity Recognition.**	
Noise (image)	**An all-purpose term for unwanted visual effects**	

	imported into a digital image by the process of image capture. Commonly caused by electrical interference in a scanner's light sensor, digital noise will most often manifest as unidentifiable specks of bright or dark colour. Noise can have deleterious effects on OCR accuracy. See also: OCR accuracy.	
OAI	**Abbreviation for Open Archives Initiative.**	
OCLC	**Abbreviation for the Online Computer Library Center.**	
OCR	**Abbreviation for Optical Character Recognition.**	
OCR accuracy	**As most commonly used, the term OCR accuracy refers to the number of correctly recognised characters/words in relation to the total number of characters/words in a document. OCR accuracy is assessed by comparing OCR results with a document's Ground truth. See also: Ground truth; Noise (image).**	
OCR engine	**An OCR engine is a system/ software that supports OCR processing.**	
Online Computer Library Center (OCLC)	**The Online Computer Library Center is a nonprofit, membership, computer library service and research organisation dedicated to the public purposes of furthering access to the world's information and reducing the rate of rise of library costs.**	OCLC [http://www.oclc.org/]
Open Archives Initiative (OAI)	**The Open Archives Initiative is the maintainer of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).**	Open Archives Initiative [http://www.openarchives.org/]
Optical Character Recognition (OCR)	**Optical character recognition is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured**	Wikipedia [http://en.wikipedia.org/wiki/Optical_character_recognition]

	by a scanner) into machine-editable text.	
Optimisation (OCR)	**As it pertains to digital images, optimisation is any technical process undertaken to prepare an image for final delivery to users. Within IMPACT, optimisation refers to processes used to enhance the OCR accuracy of a digital image.**	
Pixel	**Short for picture element; the smallest unit that makes up an image. Each pixel can represent a number of different shades or colours, depending on how much storage space is allocated for it. Pixels are also used to measure image size and resolution.**	
Pixels per inch (PPI)	**A measurement used to describe both the spatial resolution of a digital image and the physical size of an image printed from it. An image that has a higher number of pixels per inch will show more detail than one which has fewer pixels per inch.**	Wikipedia [http://en.wikipedia.org/wiki/Pixels_per_inch]
Post-processing (image)	**An umbrella term covering any form of post-capture manipulation of a digital image, such as altering brightness or contrast, changing an image's colour space to match that of an output device, segmenting an image, removing unwanted detail from an image, etc. The post-processing procedures in IMPACT all aim to improve the final OCR result of a text-based digital resource.**	
PREMIS	**Abbreviation for Preservation Metadata: Implementation Strategies, a core preservation-metadata set, supported by a data dictionary, consisting of a set of standardised terms and processes that address the provenance and characteristics of a digital item, the changes that may have been made to it**	PREMIS Schema 2.0 [http://www.loc.gov/standards/premis/premis.xsd], PREMIS Data Dictionary 2.0 [http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf]

	in the course of preservation, the technical environments in which the digital item will operate, and the intellectual ownership of the item. See also: Metadata.	
PRONOM	**PRONOM is an online information system created by the UK National Archives about data file formats and their supporting software products. PRONOM holds information about file formats and the software products which can process (i.e., read, write to and identify) each format. Information related to the file formats, such as documentation about them, their compression types, character encoding schemes and intellectual property rights is also held.**	PRONOM Technical Registry [http://www.nationalarchives.gov.uk/help/PRONOM/faq.htm#faq1]
PURL	**Abbreviation of Persistent Uniform Resource Locator. A uniform resource locator (URL) is a unique alphanumeric string that specifies where in a digital space an identified resource is located and the mechanism for retrieving it. A Persistent URL differs by not locating the resource directly, but by pointing the query to a more stable resource (e.g. a web domain rather than a web page) before redirecting to the requested item. PURLs were created as a means of managing the apparently arbitrary disappearance of content from digital name spaces.**	
QA	**Abbreviation for Quality Assurance.**	
Quality Assurance (QA)	**Quality assurance refers to planned and systematic production processes that provide confidence in a product's suitability for its intended purpose.**	Wikipedia [http://en.wikipedia.org/wiki/Quality_assurance]

RDF	**Abbreviation for Resource Description Framework.**	
Relax NG	**RELAX NG (REgular LAnguage for XML Next Generation) is a simple and popular schema language forXML. It offers a compact, non-XML syntax and some other advantages over DTD. See also: DTD.**	Relax NG Homepagehttp://relaxng.org/ [http://relaxng.org/]
Resource Description Framework (RDF)	**The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications, originally designed as a metadata data model, which has come to be used as a general method of modelling information through a variety of syntax formats. The purpose of RDF is to provide an encoding and interpretation mechanism to online resources so that they are described in a way that particular, defined software can understand. See also: Schema; Metadata.**	RDF Schema [http://www.w3.org/TR/rdf-schema/]
RGB	**The abbreviation RGB corresponds to the additive primary colours Red, Green and Blue from which the colour space for representing images in electronic systems is composed. See also: CMYK.**	sRGB Standard [http://www.w3.org/Graphics/Color/sRGB], Comparison of sRGB vs. Adobe RGB 1998 [http://www.cambridgeincolour.com/tutorials/sRGB-AdobeRGB1998.htm]
Sampling scheme	**Any method of quality assurance that groups together under consistent determining criteria a sample of the products of a particular process, taking the sample to be qualitatively representative of the whole. See also: Batch sample; Quality assurance (QA).**	
Schema	**A description of the structure and rules an xml document must satisfy in order to be considered a valid example of the type. A schema must include the formal declaration of the elements that make up a**	

	document. Examples include the DTD and RDF schemas. See also: DTD; RDF; xml.	
SDK	**Abbreviation for Software development kit.**	
Segmentation (image processing)	**The process of partitioning a digital image into multiple regions (defined as sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is easier for a machine to analyse. Image segmentation is typically used to locate objects and boundaries - such as lines and curves - within images.**	
Show-through	**An artefact of printing whereby the ink on one side of a page is clearly visible on the obverse side. Common in pre-20th century books printed on very thin paper, show-through can have severe effects on OCR accuracy. See also: OCR accuracy; Bleed-through.**	
Software development kit	**A software development kit (SDK or "devkit") is typically a set of development tools that allows a software engineer to create applications for a certain software package, software framework, hardware platform, computer system, video game console, operating system, or similar platform.**	Wikipedia [http://en.wikipedia.org/wiki/Software_development_kit]
SRU	**Abbreviation for Search/Retrieve via URL, a search protocol for queries based on CQL (Common Query Language).**	Library of Congress: SRU Standards [http://www.loc.gov/standards/sru/], Common Query Language [http://www.loc.gov/standards/sru/specs/cql.html]
Stakeholder	**A person or organisation with a definable and legitimate interest in a given situation, action or enterprise.**	
Surrogate	**A digital image derived from the archival (or master) image. Usually not as high-resolution as the archival image, surrogates are usually deployed as output images on display**	Australian National Library [http://archive.amol.org.au/capture/course/glossary.html#s]

	devices and printers. See also: Master image; Thumbnail.	
TEI	**Abbreviation for Text Encoding Initiative.**	
TEL	**Abbreviation for The European Library. TEL is a free service that offers access to digital and bibliographical resources of the 48 national libraries of Europe in 35 languages.**	The European Library [http://www.theeuropeanlibrary.org/portal/index.html]
Text Encoding Initiative (TEI)	**The Text Encoding Initiative is a consortium of institutions and research projects which collectively maintains and develops an XML standard for the representation of texts in digital form.**	TEI Consortium [http://www.tei-c.org/index.xml], TEI P5 Guidelines [http://www.tei-c.org/Guidelines/P5/]
Text Recognition	**An umbrella term for the techniques and processes used to make text readable by machines. See also: OCR; Image enhancement; Optimisation; Post-processing.**	
Text retrieval	**Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text.**	Wikipedia [http://en.wikipedia.org/wiki/Text_retrieval]
Thumbnail	**A form of surrogate image created for quick user reference in a digital space. Thumbnails will usually be much smaller than the original image and in a much lower resolution. See also: Surrogate; Master image.**	
Tool	**The term tool is used in IMPACT to refer to the software developed by the project, as well as the calculators and documents that enable and support decision-making in the field of mass digitisation of historical text-based material.**	
Typeface	**In typography, a typeface is a set of one or more fonts, in one or more sizes, designed with stylistic unity, each comprising a coordinated set of glyphs (characters). See also: Font.**	Wikipedia [file:///C:/Daten/Arbeit/en.wikipedia.org/wiki/Typeface]

Unicode	**In computing, Unicode is an industry standard allowing computers to consistently represent and manipulate text expressed in most of the world's writing systems.**	Unicode Consortium [http://www.unicode.org/], Wikipedia [http://en.wikipedia.org/wiki/Unicode]
URI	**An abbreviation of Uniform Resource Identifier, which identifies the name and address of information on a shared, browsable network. A URI usually identifies the application used to access an information resource, the machine the resource is located on, and the file name of the resource. A Uniform Resource Locator (URL) is a type of URI. See also: URL; PURL.**	
URL	**An abbreviation of Uniform Resource Locator, which is the address of a file or content on a shared, browsable network such as the internet.**	
Warping of paper	**Due to age and the relative humidity of its storage conditions, the pages of books and other bound volumes will often be warped rather than flat. Warped pages have a high negative effect on OCR accuracy, but may be mitigated by electronic dewarping techniques. See also: Dewarping.**	
Workflow	**A workflow is a repeatable pattern of activity enabled by a systematic organisation of resources, defined roles and mass, energy and information flows, into a work process that can be documented and learned.**	Wikipedia [http://en.wikipedia.org/wiki/Workflow]
XML	**Abbreviation for Extensible Markup Language.**	
XML Schema	**XML Schema, published as a World Wide Web Consortium (W3C) recommendation in May 2001, is one of several XML schema languages. See also: Schema; DTD; RDF.**	W3C Recommendation [http://www.w3.org/TR/xmlschema-0/]

Yellowed paper	Due to age, its chemical makeup or environmental storage factors, paper may change its colour over time. This can cause minor negative effects to the OCR process.