

*conversion of newspapers
to digital objects,
digital data preservation,
and other interesting things*

Frederick Zarndt
Chair, IFLA Newspapers Section

frederick@frederickzarndt.com

70 Barnard

DESERET NEWS.

BY W. RICHARDS.

G. S. L. CITY, DESERET, JULY 1, 1855.

VOL. I. -- NO. 5.

LAT. 40° 45' 44" LON. 111° 26' 34"

DESERET NEWS,
PUBLISHED EVERY SATURDAY.

TERMS.

6 Months, \$2.50; in advance.
Single copy, 15 cents.
One doz., 12 1-2 cents each.

ADVERTISING.

Per square, 14 lines, \$1.50.
Succeeding insertions, 50 cents each.
\$1.00 for a half square.

EMIGRANTS AND TRAVELLERS.

Can have their names, place of residence, time of arrival and departure, inserted in the NEWS and a copy mailed to their friends for 25 cents.

Ten cents for insertion, without the paper.

Companies of 20 and upwards, entered at once, 20 cents each.

Any additional information, 10 cents per line.

NEWS.

DELIVERED at the Post Office, which will be open, each Sabbath from 12 to 1 o'clock P. M.

AGENTS.

ANSON CALL, North Canyon.
DANIEL MILLER, N Cottonwood.
ISAAC CLARK, Weber County.
JOEL H. JOHNSON, Mill Creek.
WM. CROSBY, Cottonwood.
ISAAC HIGBEE, Utah.
PHINEAS RICHARDS, San Pete.
EZRA T. BENSON, Tooele.

BISHOP HOLLADAY, and all the acting Bishops in the City.

Unless subscribers advise us to the contrary, we shall send their papers, to our agent nearest their residence.

JOB PRINTING. AT THE NEWS OFFICE.

U. S. MAIL.

Is expected to leave for the States, about the 27th, of July.

POSTAGE.

Single letters to any part of the States, 40 cents.

Papers mailed from our office, can pay postage on receipt, in the States.

GRAND CONCERT.

Our friends, fellow citizens and emigrants are respectfully informed that there will be a Grand Concert in the Bowers, on the evening of Wednesday the 24th, inst.

As the people love amusement, we design to gratify them, with a series of Comic pieces and songs, most of which will be entirely new in this Valley and some original, got up *expressly* for the occasion. For particulars see hand bills.

Admittance by Tickets, which can be had at the Tithing and Post Office, 25cts each.

WM. CLAYTON.

C. P. T.

G. S. L. City. July 10th, 1855.

NOTICE.

ALL persons, that have branded cattle or horses, residing in the valley, are hereby notified, that when they trade the same with emigrants, or others, the law requires them, to reverse the brand, either above or below the original, which guarantees the vend,

MISS NANNY, A NOTED PETRICKER.

Though Not Hard to PLEASE.

I don't like a man that's fat—

A man that's lean is worse than that;

Nor do I like a man that's tall—

A man that's little, is worse than all;

Nor do I like a man that's fair—

A man that's dark I cannot bear;

A young man is a constant pest—

An old man would my hours incest;

A man of sense I could not rule;

And yet I could not love a fool;

A sinner man I will not take—

A drunkard and my heart would break.

Man's face but I should hate to see.

But worse than all a GREAT COATER.

All these I most sincerely hate,

And yet I love the MARRED STATE.

DAMAGES. Two gardens were destroyed on Tuesday night by emigrants cattle, which cost them \$74. Our Marshall suggests that it would be wisdom for the emigrants to camp further from the city, thereby saving their money, and leaving the vegetables to grow.

SAN PETE. Several Brothron arrived from San Pete on Tuesday bringing 34 M. Shingles, and report all well; crops late but prosperous.

We are informed that Estell & Co. of Weston, Mo. are running a mail from Mo. to Pacific Springs, accommodating all travellers, on the route at 50 cts., per letter.

The Council of Health meet on Wednesday; advice gratis, from 3 to 4 P. M.

A drove of cows passed our office on Tuesday *en route* for California.

We understand the main body of emigration is about 2 weeks back of this point.

A person who undertakes to raise himself by scandalizing others, might as well sit down on a wheelbarrow and undertake to wheel himself.

*why digitize
newspapers?*













Photo by DAVID ILIFF. License: CC-BY-SA 3.0

reading rooms by the numbers

			Monthly average		
			Visitors	Requests for Newspapers	
		Population	Reading Room	Microform	Print
	Australia	22,876,000	5,130	345	240
	France	65,350,000	3,000	2,000	1,000
	Netherlands	16,847,000	NA	NA	NA
	New Zealand	4,414,000	NA	NA	NA
	Norway	4,985,000	600	400	NA
	Singapore	5,184,000	NA	300	NA
	UK	62,262,000	2,000	6,900	4,816
	USA	313,292,000	NA	NA	NA













digitised newspapers by the numbers

		Monthly average			
		Digitised Historical Newspapers			
	Population	Unique Visitors	Genealogist	Other	User Age
	22,876,000	150,000	50%	50%	>55
	37,692,000	12,800	65%	35%	>50
	5,405,000	NA	NA	NA	?
	65,350,000	22,000	NA	NA	?
	16,847,000	50,000	NA	NA	?
	4,414,000	83,333	50%	NA	>50
	4,985,000	1,500	NA	NA	?
	5,184,000	12,400	NA	NA	?
	62,262,000	NA	NA	NA	?
	313,292,000	NA	NA	NA	?

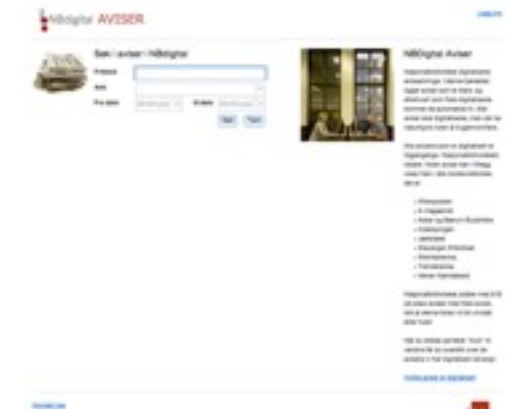












physical versus digital



		Monthly average	
		Requests for Newspapers	Digitised Historical Newspapers
	Population	Paper + Microform	Unique Visitors
	22,876,000	585	150,000
	37,692,000	NA	12,800
	5,405,000	NA	NA
	65,350,000	3,000	22,000
	16,847,000	NA	50,000
	4,414,000	NA	83,333
	4,985,000	400	1,500
	5,184,000	300	12,400
	62,262,000	11,716	NA
	313,292,000	NA	NA

more numbers!



				Monthly average				
	Collection			Digitised Historical Newspapers				
	Population	Name	~Size [pages]	Unique Visitors	Genealogist	Other	Lines Corrected	User Age
	22,876,000	Trove	5,000,000	150,000	50%	50%	220,000	>55
	37,692,000	CDNC	495,000	12,800	65%	35%	31,000	>50
	5,405,000	Historical Newspaper Library	2,000,000	NA	NA	NA	NA	?
	65,350,000	Gallica	2,200,000	22,000	NA	NA	NA	?
	16,847,000	Historische Kranten	5,000,000	50,000	NA	NA	NA	?
	4,414,000	Papers Past	2,213,000	83,333	50%	NA	NA	>50
	4,985,000	NBDigital Aviser	8,100,000	1,500	NA	NA	NA	?
	5,184,000	Newspaper SG	2,400,000	12,400	NA	NA	NA	?
	62,262,000	British Newspaper Archive	4,880,000	NA	NA	NA	NA	?
	313,292,000	Chronicling America	4,100,000	NA	NA	NA	NA	?











what is Alexa?

- Alexa collects and analyzes Internet data for purposes of web analytics. Web analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing web usage. Alexa is now a subsidiary of Amazon.
- Alexa was founded in 1996 by Brewster Kahle (Internet Archive) and Bruce Gilliat.
- Alexa operations includes archiving of webpages as they are crawled. This database served as the basis for the creation of the Internet Archive accessible through the Wayback Machine.
- Alexa continually crawls all publicly-available websites to create a series of snapshots of the web.
- Alexa gathers information from a variety of sources to provide key statistics about each site on the web, for example, Traffic **Rank**, the number of **PageViews**, and site **Speed**, **Bounce Rate**, etc. This information is derived from Alexa toolbar users (~6,000,000 worldwide).









definitions

- A **PageView** is a request for a file whose type is defined as a page.
- A **Unique Visitor** is a uniquely identified client generating requests on the web server or viewing pages within a defined time period (i.e. day, week or month). A Unique Visitor counts once within the timescale.
- A **Visit** is a series of page requests from the same uniquely identified client with a time of no more than 30 minutes between each page request.
- **Bounce Rate** is the percentage of visits where the visitor enters and exits at the same page without visiting any other pages on the site in between.
- World | Country **Rank** is a function of the average daily unique visits and the number of unique pages requested.


Alexa ranking world view

	Alexa 3 month trailing averages 2-Apr-2012		
	Population	Website	World rank [Lo is good]
	313,292,000	http://www.loc.gov/index.html/	3,122
	22,876,000	http://trove.nla.gov.au/	16,700
	65,350,000	http://www.bnf.fr/	17,096
	62,262,000	http://www.bl.uk/	27,079
	4,414,000	http://www.natlib.govt.nz/	123,976
	62,262,000	http://www.britishnewspaperarchive.co.uk/	155,259
	16,847,000	http://www.kb.nl/	155,363
	5,184,000	http://www.nl.sg/	156,610
	4,985,000	http://www.nb.no/	189,940
	5,405,000	http://www.nationallibrary.fi/	3,212,803

Alexa ranking country view











Alexa 3 month trailing averages 2-Apr-2012				
	Population	Website	World rank [Lo is good]	Country rank [Lo is good]
	5,405,000	http://www.nationallibrary.fi/	3,212,803	199
	22,876,000	http://www.nla.gov.au/	16,700	375
	4,414,000	http://www.natlib.govt.nz/	123,976	515
	65,350,000	http://www.bnf.fr/	17,096	727
	4,985,000	http://www.nb.no/	189,940	891
	313,292,000	http://www.loc.gov/index.html/	3,122	1,011
	5,184,000	http://www.nl.sg/	156,610	1,208
	62,262,000	http://www.bl.uk/	27,079	2,245
	16,847,000	http://www.kb.nl/	155,363	3,450
	62,262,000	http://www.britishnewspaperarchive.co.uk/	155,259	15,692

where visitors go

Alexa 3 month trailing averages 2-Apr-2012					
	Population	World rank [Lo is good]	Country rank [Lo is good]	Where visitors go [sub-domain]	
	5,405,000	3,212,803	199	NA	NA
	22,876,000	16,700	375	http://trove.nla.gov.au/	57.2%
	4,414,000	123,976	515	http://paperspast.natlib.govt.nz/	50.9%
	65,350,000	17,096	727	http://gallica.bnf.fr/	52.0%
	4,985,000	189,940	891	NA	NA
	313,292,000	3,122	1,011	http://chroniclingamerica.loc.gov/	4.8%
	5,184,000	156,610	1,208	http://newspapers.nl.sg/	28.0%
	62,262,000	27,079	2,245	http://newspapers11.bl.uk/blcs/	2.5%
	16,847,000	155,363	3,450	http://kranten.kb.nl/	22.4%
	62,262,000	155,259	15,692	NA	NA













lots of numbers

(sorted by time on site)

		Alexa 3 month trailing averages 2-Apr-2012				
	Website	Speed [Hi is good]	Bounce rate [Lo is good]	Reputation [Hi is good]	Page views per visitor [Hi is good]	Time on site [Hi is good]
	http://www.britishnewspaperarchive.co.uk/	51%	28%	485	13.0	11m 40s
	http://www.bnf.fr/	71%	35%	13,744	14.9	8m 30s
	http://www.natlib.govt.nz/	96%	44%	2,480	5.3	6m 49s
	http://trove.nla.gov.au/	42%	55%	9,514	5.4	4m 52s
	http://www.loc.gov/index.html/	67%	51%	91,331	5.3	3m 55s
	http://www.kb.nl/	89%	54%	3,295	5.0	3m 42s
	http://www.bl.uk/	54%	52%	16,191	3.8	3m 2s
	http://www.nb.no/	59%	47%	1,579	3.0	2m 57s
	http://www.nationallibrary.fi/	NA	54%	199	3.1	2m 6s
	http://www.nl.sg/	72%	65%	802	2.0	2m 4s

even more numbers

(sorted by time on site)

		Alexa 3 month trailing averages 2-Apr-2012				
	Website	Speed [Hi is good]	Bounce rate [Lo is good]	Reputation [Hi is good]	Page views per visitor [Hi is good]	Time on site [Hi is good]
	http://www.ancestry.com/	32%	24%	20,055	29.9	23m 54s
	http://www.familysearch.org/	50%	18%	9,832	15.8	16m 19s
	http://www.britishnewspaperarchive.co.uk/	51%	28%	485	13.0	11m 40s
	http://www.bnf.fr/	71%	35%	13,744	14.9	8m 30s
	http://www.natlib.govt.nz/	96%	44%	2,480	5.3	6m 49s
	http://trove.nla.gov.au/	42%	55%	9,514	5.4	4m 52s
	http://www.loc.gov/index.html/	67%	51%	91,331	5.3	3m 55s
	http://www.kb.nl/	89%	54%	3,295	5.0	3m 42s
	http://www.bl.uk/	54%	52%	16,191	3.8	3m 2s
	http://www.nb.no/	59%	47%	1,579	3.0	2m 57s
	http://www.nationallibrary.fi/	NA	54%	199	3.1	2m 6s
	http://www.nl.sg/	72%	65%	802	2.0	2m 4s

why digitize newspaper collections?



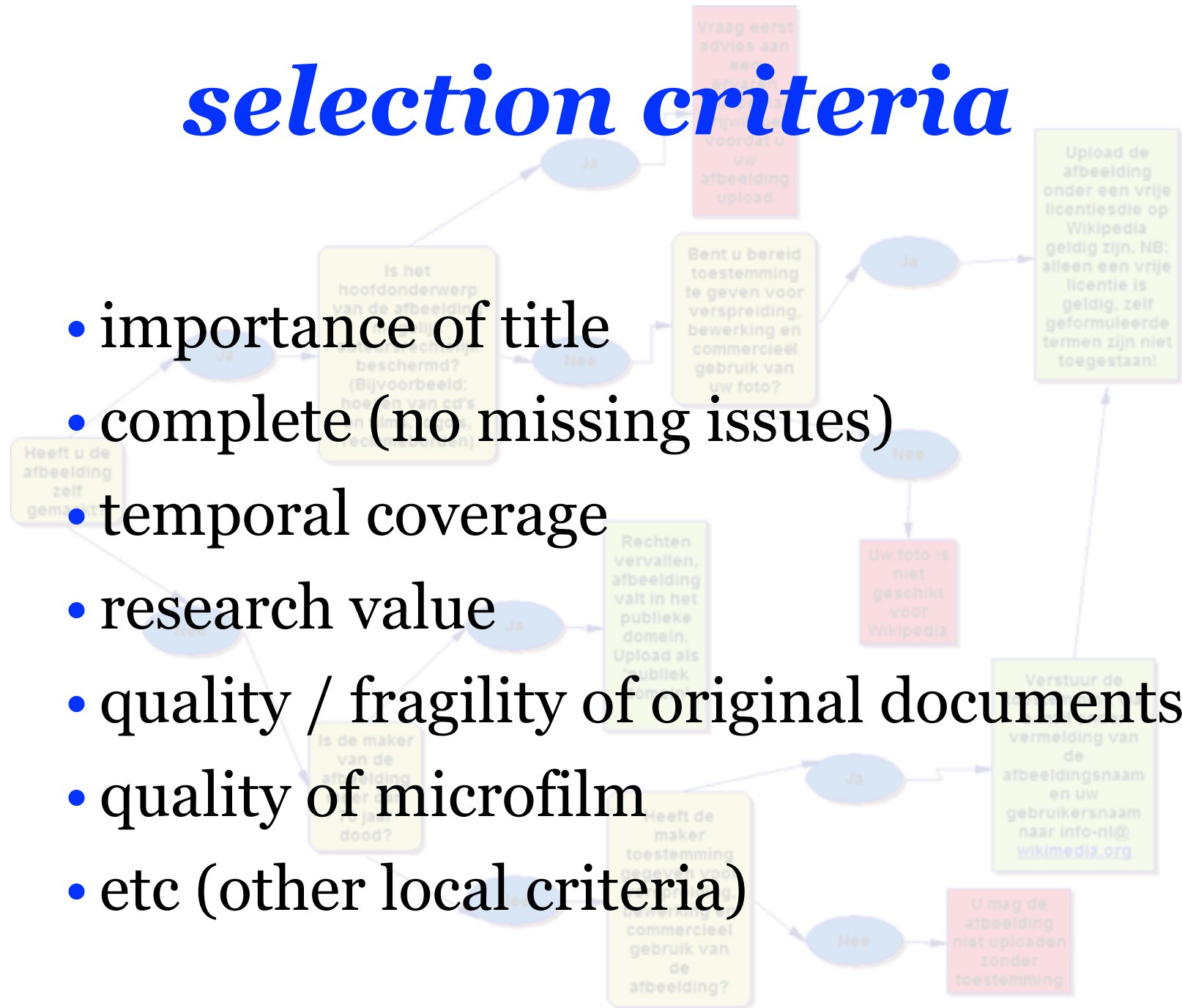
*digital newspapers enable broader,
easier, and faster access*



considerations in newspaper digitization

selection criteria

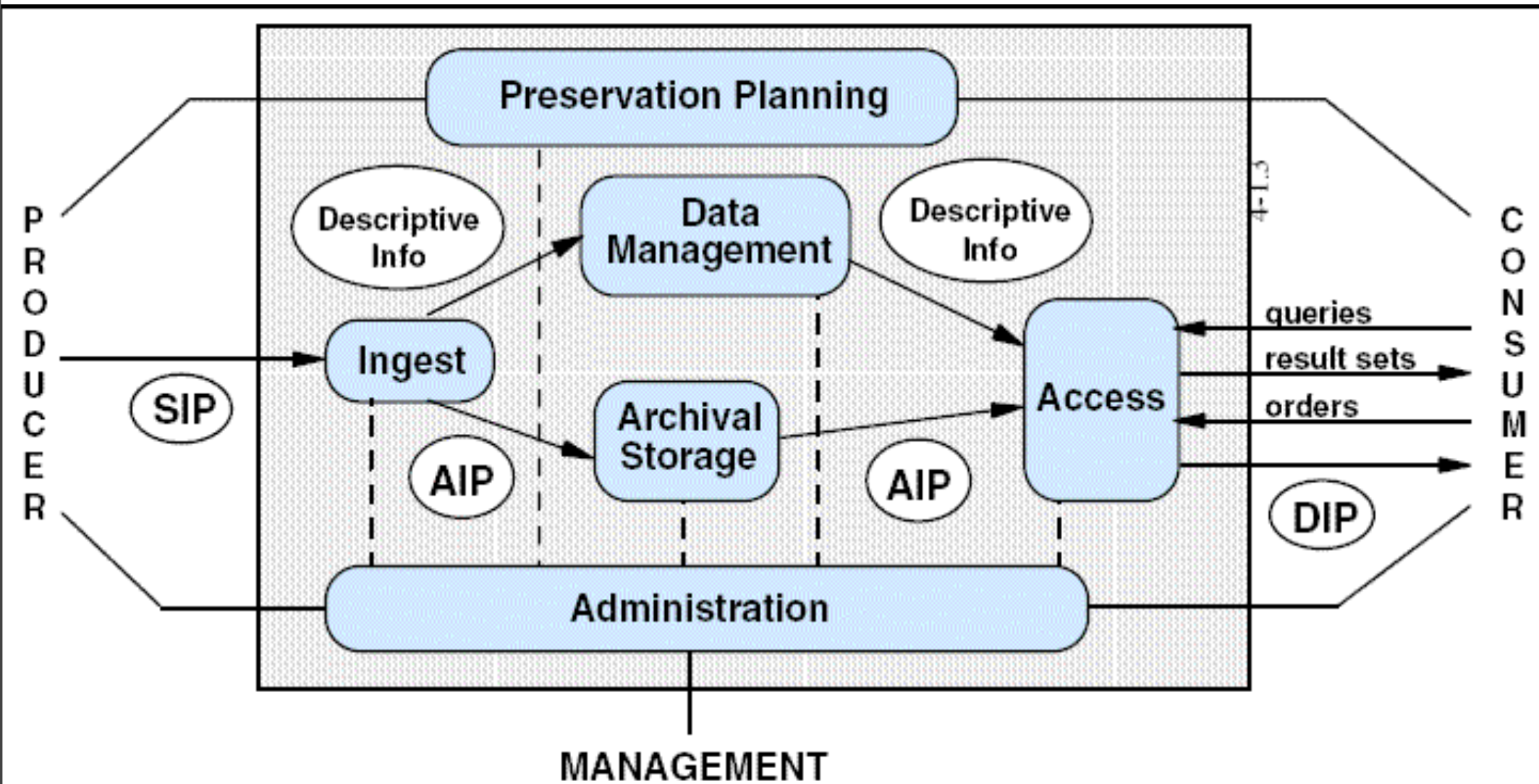
- importance of title
- complete (no missing issues)
- temporal coverage
- research value
- quality / fragility of original documents
- quality of microfilm
- etc (other local criteria)



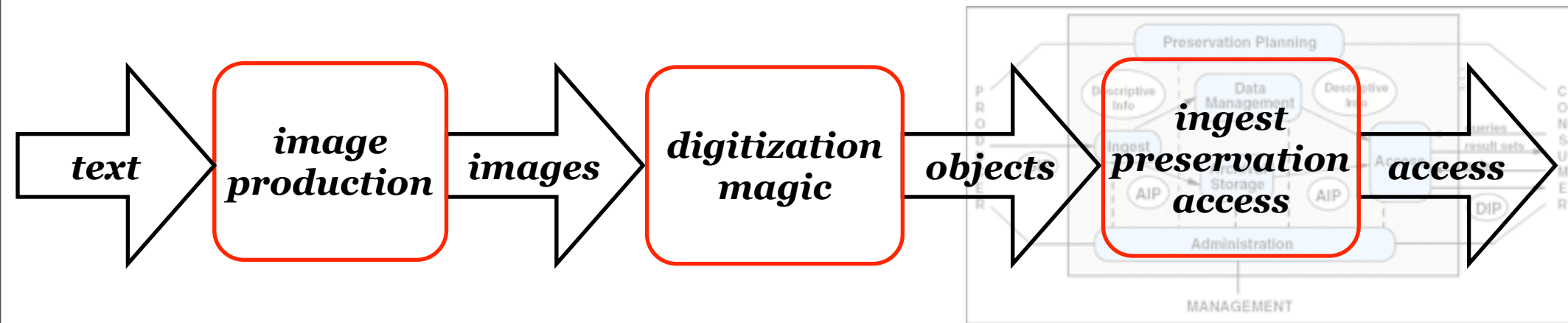
page-level versus article-level newspaper digitization

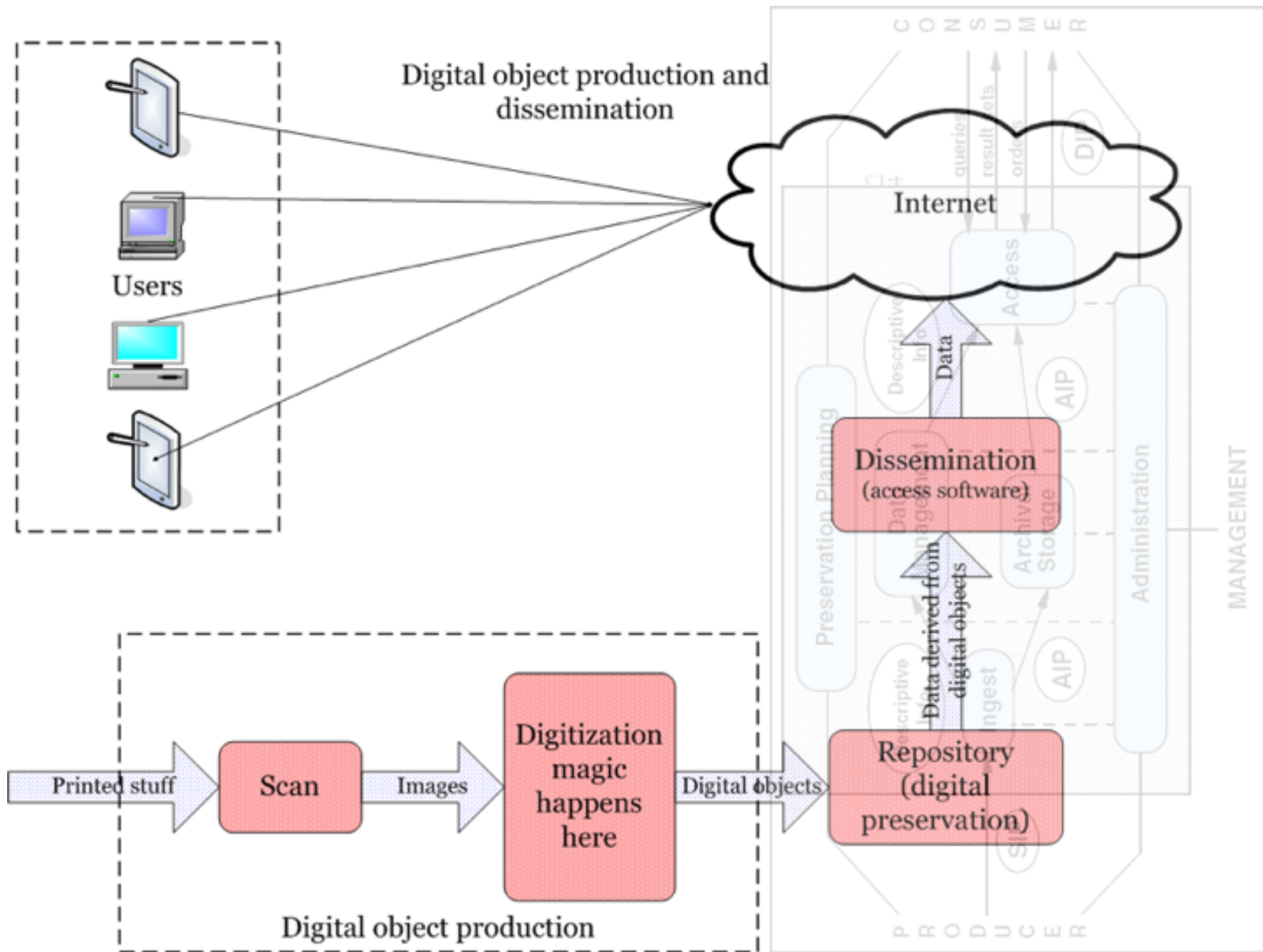
	cost	production difficulty	copyright management	usability	accessibility
page-level	\$	easy	usually simple	low	good
article-level	\$\$\$	hard	usually complex	excellent	excellent

preservation, access, administration *Open Archival Information System* *(OAIS) reference model*



the digitization process





standard file formats

- image file formats
 - TIFF
 - JPEG2000
 - JPEG
 - GIF
- text file formats
 - PDF, PDF/A, PDF/A-1b, PDF/A-1a
 - TEI XML
 - HTML
 - plain text
 - NITF / NewsML
- metadata
 - METS
 - MODS / PREMIS / ALTO / MIX ...

image decisions

- image production source materials
 - original documents: better quality, more expensive
 - microfiche: poorer quality, less expensive, microfiche quality varies
- bit depth
 - black-and-white (bitonal)
 - greyscale
 - color
- resolution
- compression
 - no compression
 - lossless (reversible)
 - lossy (irreversible)
- image metadata

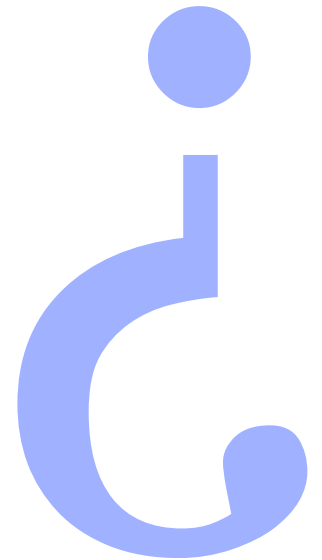
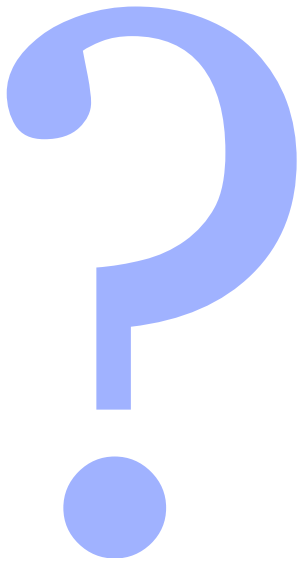


image format comparison

	compression	bit depth	metadata	color management	mime type	patent	1st public release
JBIG (.jbig, .jbg)	lossless	1-bit		no		no	2000?
JPEG (.jpg, .jpeg)	lossy, DCT, RLE, Huffman	8-bit 12-bit 24-bit	yes	yes	image/jpeg public.jpeg	no	1992
JPEG2000 (.jp2)	many lossless and lossy compression algorithms	8-bit 16-bit color to 48 bits	yes	yes	image/jp2 public.jpeg2000	yes but part 1 is patent free	2000
TIFF (.tiff, .tif)	none LZW RLE ZIP Other	1, 2, 4, 8, 16, 24, 32 bits	yes	yes	image/tiff public.tiff	no	1986

digital library standards

- METS XML for descriptive, structural, technical, and administrative metadata
- descriptive metadata
 - Metadata Object Description Standard (MODS)
selected metadata from MARC
 - Dublin Core fundamental group of text elements for describing and cataloging
- technical metadata
 - ALTO for OCR text
 - PREMIS for digital preservation
 - MIX and ANSI/NISO Z39.87 for images

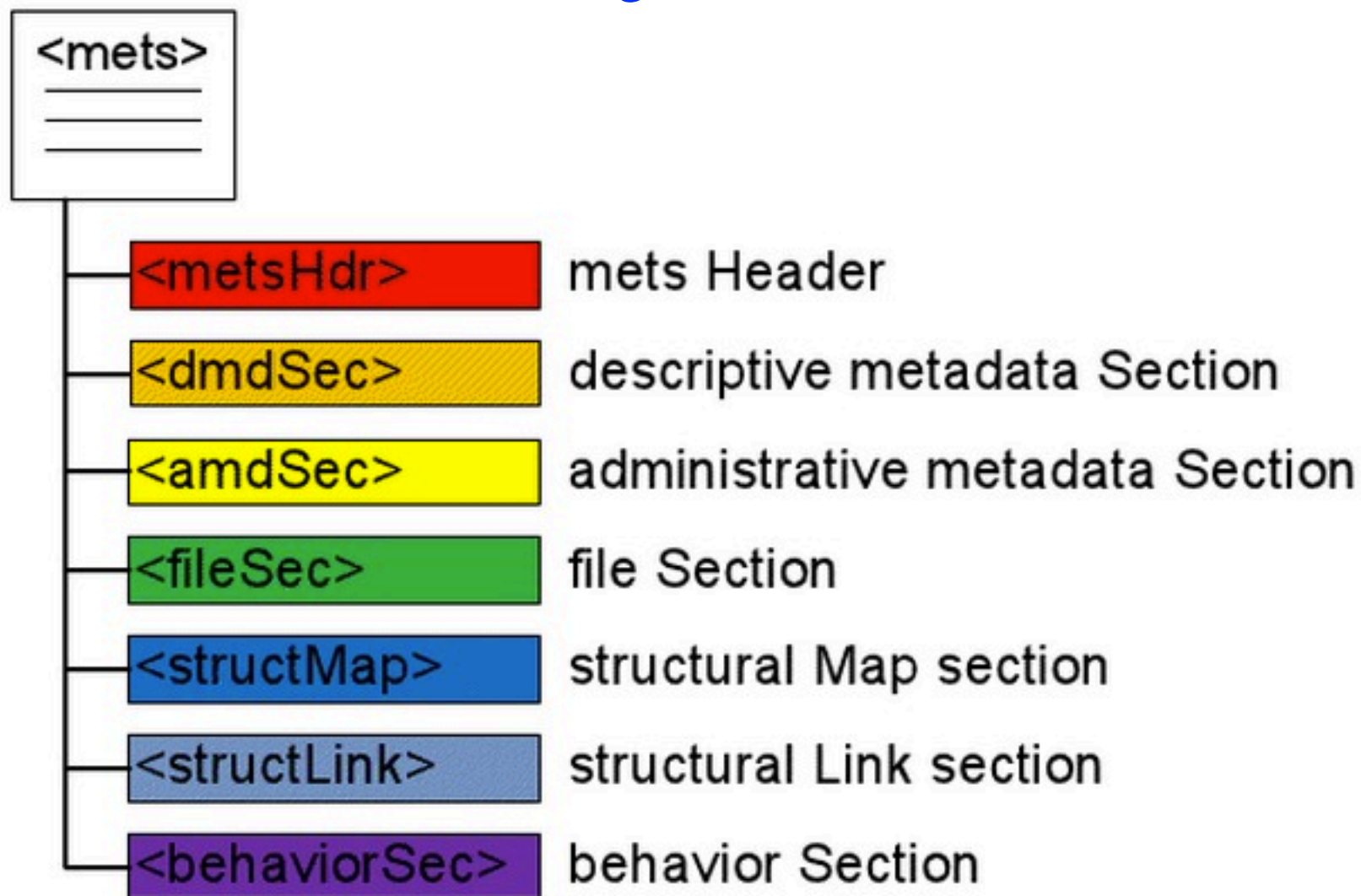




Metadata Encoding and Transmission Standard

- METS is a XML standard for encoding descriptive, administrative, and structural metadata about objects within a digital library
- METS files consist of 7 (optional) sections: header, descriptive, administrative, file map, structural map, structural link, and behavior
- METS profiles describe a class of METS documents in sufficient detail to provide both document authors and programmers the guidance to create and process METS documents conforming with a particular profile
- current version 1.9.1
- administered by METS editorial board (international group of volunteers)
- standards hosted by Library of Congress at <http://www.loc.gov/standards/mets/>

METS file structure



Metadata Object Description Schema

- MODS is an XML schema for a bibliographic element set that may be used for library applications. Derivative of MARC 21 bibliographic format. Includes a subset of MARC fields, using language-based tags rather than numeric ones
- Subset of MARC 21
- Mappings exist between MODS and MARC, Dublin Core, and RDA (conversion tools exist)
- May be used in conjunction with METS XML
- current version 3.4
- administered by Library of Congress Network Development and MARC Standards Office with help from interested users
- standards hosted by Library of Congress at <http://www.loc.gov/standards/mods/>

MODS

MODS metadata in METS XML

```
<mets:dmdSec ID="issue-nla.news-issn18368190_18740425">
  <mets:mdWrap MDTYPE="MODS">
    <mets:xmlData>
      <mods:mods xmlns="http://www.loc.gov/mods/v3">
        <mods:language>
          <mods:languageTerm type="code" authority="rfc3066">en</mods:languageTerm>
        </mods:language>
        <mods:genre>newspaper issue</mods:genre>
        <mods:originInfo>
          <mods:dateIssued>18740425</mods:dateIssued>
        </mods:originInfo>
        <mods:relatedItem type="host">
          <mods:titleInfo>
            <mods:title>The Queenslander (Brisbane, Qld. : 1866-1939)</mods:title>
          </mods:titleInfo>
          <mods:genre>newspaper</mods:genre>
          <mods:identifier>ISSN18368190</mods:identifier>
          <mods:part>
            <mods:detail type="volume">
              <mods:number>IX</mods:number>
            </mods:detail>
          </mods:part>
          <mods:part>
            <mods:detail type="issue">
              <mods:number>12</mods:number>
            </mods:detail>
          </mods:part>
        </mods:relatedItem>
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
```

MODS

Dublin Core metadata

- Dublin Core is a set of vocabulary terms used to describe resources for the purposes of discovery.
- Dublin Core metadata element set is endorsed in IETF RFC 5013, ISO 15836-2009, and NISO Z39.85
- Metadata terms last updated 14-Jun-2012
- May be used in conjunction with METS XML
- Dublin Core Metadata Initiative (DCMI) is an open organization, incorporated as a public, not-for-profit company in Singapore
- Dublin Core Metadata Initiative is hosted at <http://dublincore.org/>

Analyzed Layout and Text Object

- ALTO XML provides technical metadata for describing the layout and content of physical text resources, such as pages of a book or a newspaper
- commonly used in conjunction with METS XML but may be used standalone
- current version 2.0
- administered by ALTO editorial board (international group of volunteers)
- standards hosted by Library of Congress at <http://www.loc.gov/standards/alto/>

Analyzed Layout and Text Object book

CHAPTER I. THE CONSTITUTION OF MILK.

I. PURPOSE OF MILK.

Cow's milk is given for the primary purpose of nourishing the young calf until it can seek other food in variety.

II. COMPOSITION.

One might therefore expect to find that it contains all the food elements necessary for the building up of the young animal's body. An analysis reveals the presence of water, for the young animal's body is in the largest proportion composed of water; ash for the bones; nitrogenous material in the form of casein, albumose and albumen to nourish the muscles, hair, hoofs and horns; and carbonaceous matter in the form of sugar and fat to maintain the heat of the body.

The following table will give a fair idea of the average composition of milk as delivered to a New York cheese factory; the figures being taken from Bulletin 82, December, 1894, Geneva, New York Experiment Station:

TABLE SHOWING AVERAGE MONTHLY COMPOSITION OF MILK.

Month.	Per cent Water.	Per cent Total Solids.	Per cent of Fat.	Per cent Solids—not Fat.	Per cent Nitrogen Compounds.	Per cent Casein.	Per cent Albumen.	Per cent Albumose.	Per cent Sugar, Ash, Etc.
May	87.40	12.60	3.63	8.97	3.14	2.44	0.32	0.38	5.83
June	87.53	12.47	3.55	8.92	3.07	2.35	0.29	0.43	5.85
July	87.63	12.37	3.59	8.78	3.00	2.27	0.29	0.44	5.78
August	87.51	12.49	3.78	8.71	3.05	2.32	0.31	0.42	5.66
September ...	87.33	12.67	3.75	8.92	3.10	2.41	0.34	0.35	5.82
October	86.87	13.13	4.00	9.13	3.36	2.60	0.36	0.40	5.77

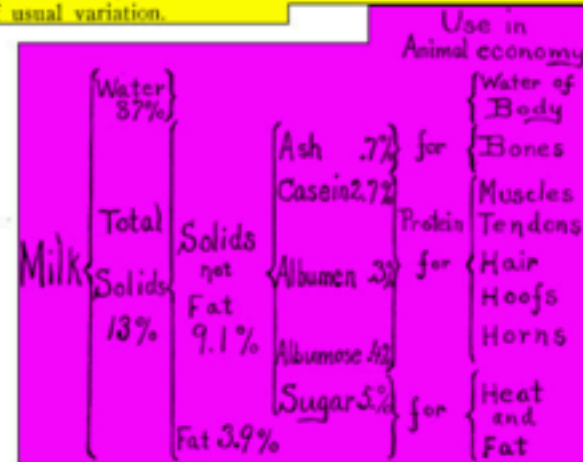
These samples were not fresh when received by the chemist, part of the albumen having been changed to albumose. It is

CHEESE MAKING.

given in the table as reported by the chemist, but the albumen and albumose may be thought of as albumen.

This table shows that the total solids in the milk varies between 12 and 13 per cent, and the fat varies between 3.5 and 4.7 per cent. These are averages for the milk in the vat at the factory. Individual cows or herds may produce milk varying considerably from these averages. In the table the sugar, ash, etc., are combined. Approximately speaking milk contains 5 per cent of milk sugar and .7 per cent ash.

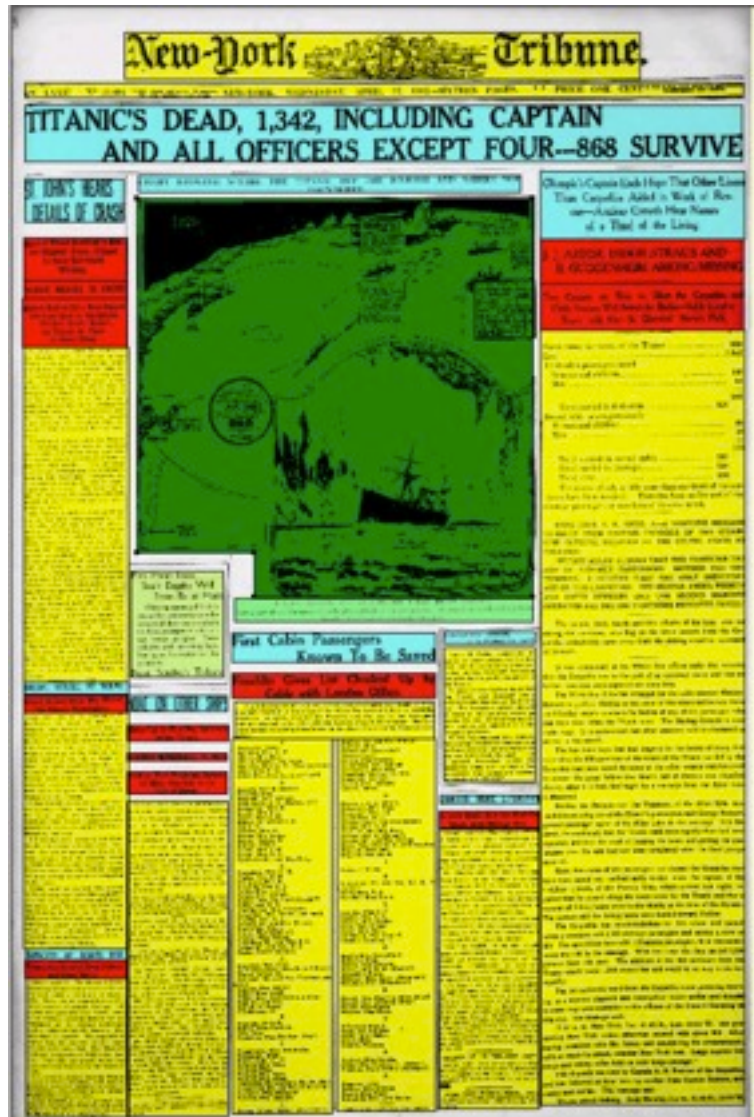
The following chart shows how the different constituents of the milk are usually grouped with an approximate relation to their use as food in the animal economy. Thousands of milk analyses are on record, but these vary some with conditions of location, etc., so that it would be difficult to give an absolutely correct average, but the figures here given are within the range of usual variation.



MAN'S USE OF MILK.

Man has diverted milk from its normal purpose (the nourishment of the calf) and uses it for a number of food products for himself. The cow normally gives enough milk in quantity and duration to nourish the calf until it can care for itself and then dries up; but by artificial means the cow has been accustomed to the habit of giving milk in larger quantities and for a

Analyzed Layout and Text Object newspaper



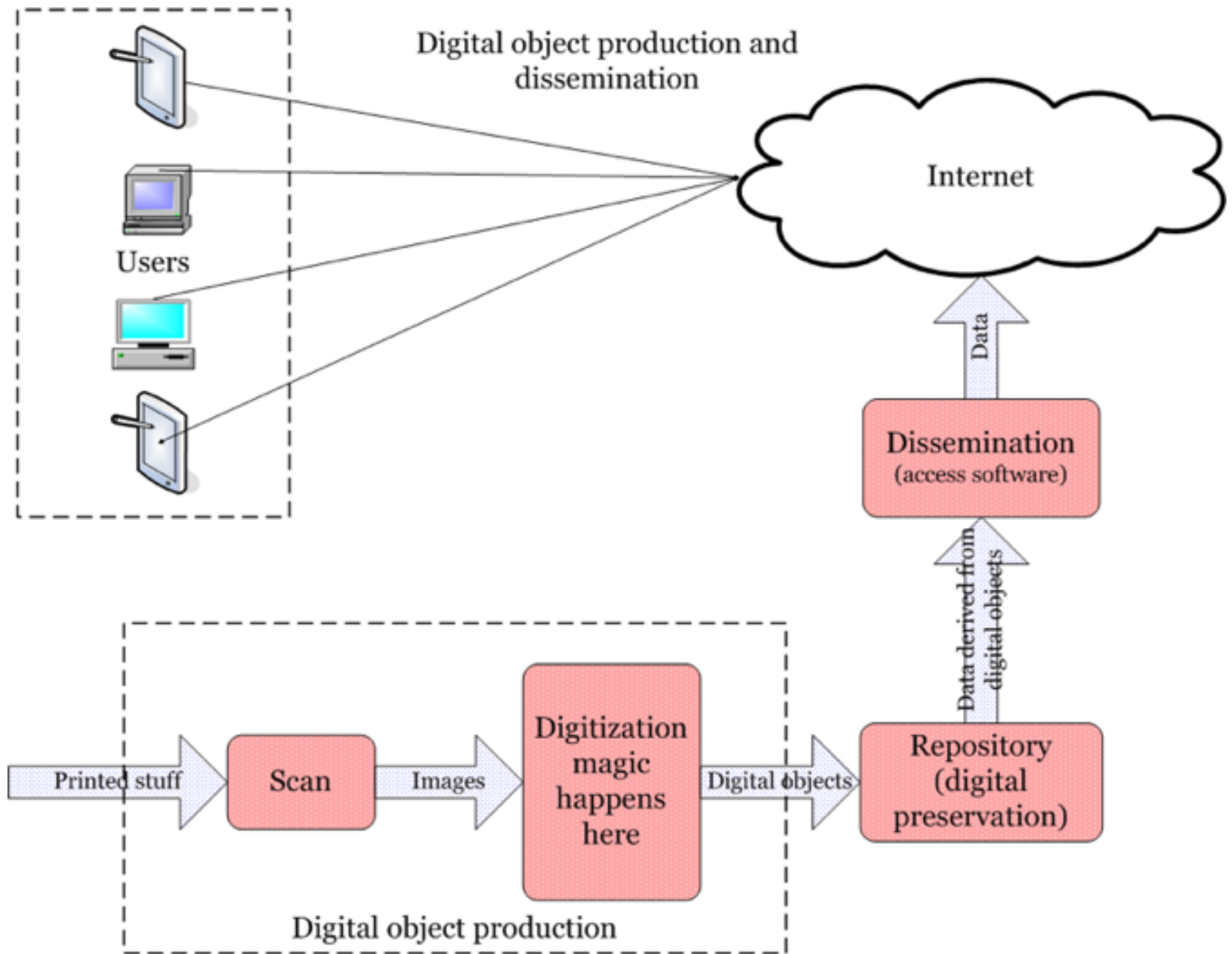


Preservation Metadata Implementation Strategies

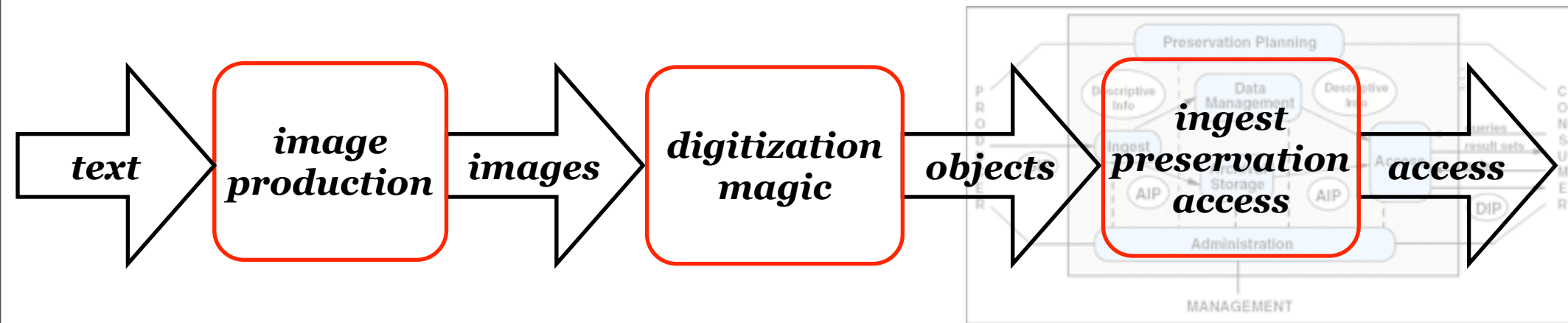
- PREMIS is a core set of implementable preservation metadata, broadly applicable across a wide range of digital preservation contexts and supported by guidelines and recommendations for creation, management, and use
- In 2003 OCLC and RLG jointly sponsored the formation of the PREMIS working group comprised of international experts in the use of metadata to support digital preservation activities
- PREMIS data dictionary current version 2.2
- May be used in conjunction with METS XML
- PREMIS tools are freely available
- PREMIS Maintenance Activity and Editorial Committee has international members from libraries and industry
- PREMIS data dictionary is hosted at <http://www.loc.gov/standards/premis/>

PREMIS data in METS file

```
<mets:amdSec>
  <mets:techMD ID="PREMISOBJECT1">
    <mets:mdWrap MDTYPE="PREMIS">
      <mets:xmlData>
        <premis:object xmlns:premis="http://www.loc.gov/standards/premis/v1">
          <premis:objectIdentifier>
            <premis:objectIdentifierType>National Library of Australia</premis:objectIdentifierType>
            <premis:objectIdentifierValue>nlaImageSeq-218-b.tif</premis:objectIdentifierValue>
          </premis:objectIdentifier>
          <premis:objectCategory>file</premis:objectCategory>
          <premis:objectCharacteristics>
            <premis:format>
              <premis:formatDesignation>
                <premis:formatName>TIFF</premis:formatName>
                <premis:formatVersion>TIFF 6.0</premis:formatVersion>
              </premis:formatDesignation>
            </premis:format>
          </premis:objectCharacteristics>
          <premis:relationship>
            <premis:relationshipType>derivation</premis:relationshipType>
            <premis:relationshipSubType>is derivative of</premis:relationshipSubType>
            <premis:relatedObjectIdentification>
              <premis:relatedObjectIdentifierType>National Library of Australia</
premis:relatedObjectIdentifierType>
              <premis:relatedObjectIdentifierValue>nlaImageSeq-218-b.tif</premis:relatedObjectIdentifierValue>
              <premis:relatedObjectSequence>0</premis:relatedObjectSequence>
            </premis:relatedObjectIdentification>
            <premis:relatedEventIdentification>
              <premis:relatedEventIdentifierType>National Library of Australia</
premis:relatedEventIdentifierType>
              <premis:relatedEventIdentifierValue>deskew-nlaImageSeq-218-b.tif</
premis:relatedEventIdentifierValue>
              <premis:relatedEventSequence>0</premis:relatedEventSequence>
            </premis:relatedEventIdentification>
          </premis:relationship>
        </premis:object>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:techMD>
</mets:amdSec>
```

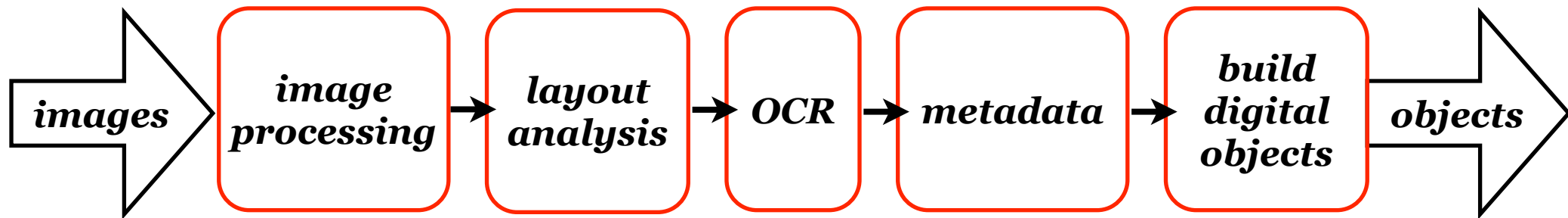
the digitization process



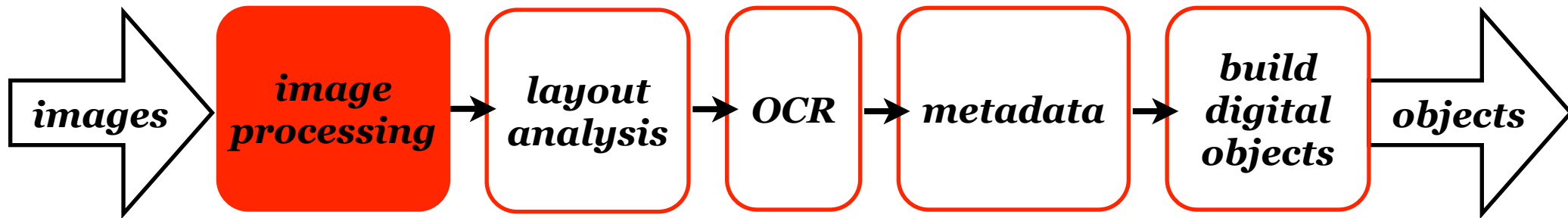
digitization magic



digitization magic

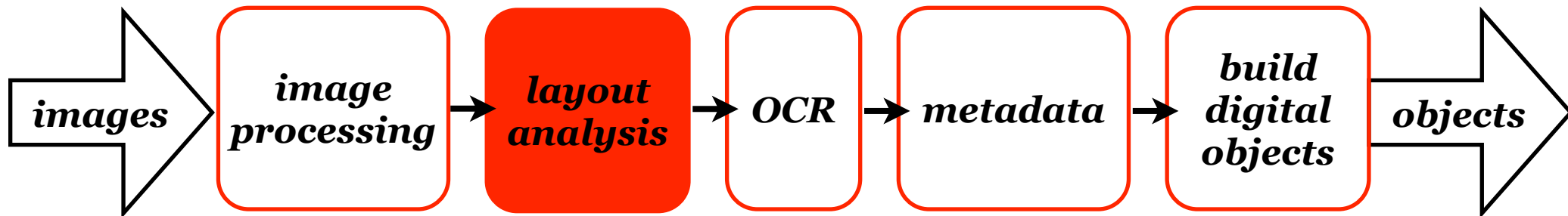


digitization magic



- crop, de-skew, split images
- apply image improvement algorithms as needed
 - sharpening filters
 - local adaptive thresholding
 - remove text bleed-thru
 - etc
- create master images
- create working images

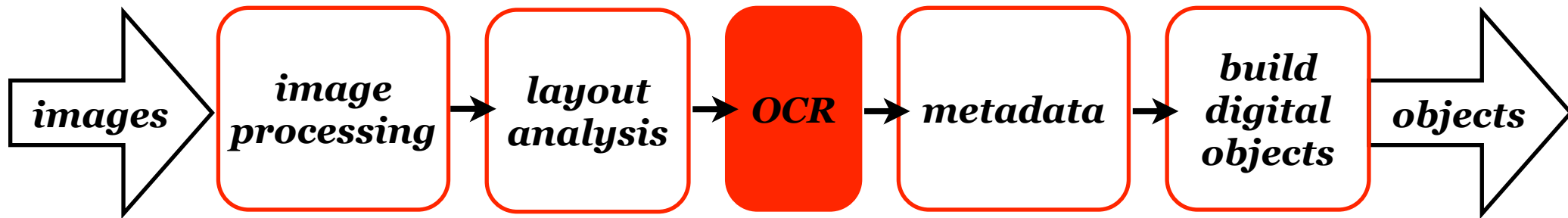
digitization magic



- analyze layout of text image
- estimate font types and sizes
- calculate coordinates of text blocks
- determine layout object types (text, illustration, headline, etc)

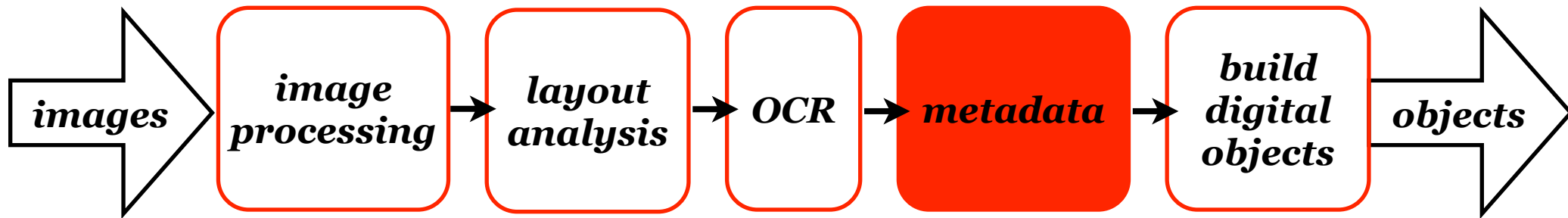


digitization magic



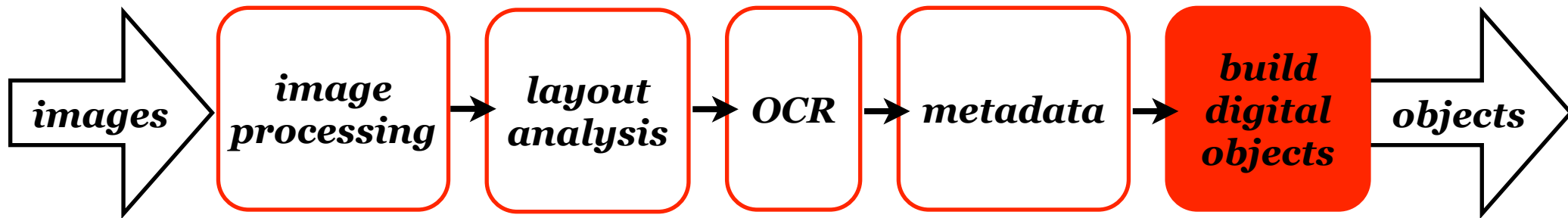
- perform optical character recognition (OCR)
- calculate word and character coordinates
- calculate word and character confidences
- apply language dictionaries
- correct OCR text (optional)

digitization magic



- populate metadata fields
- verify / correct page numbers
- verify / correct document structure

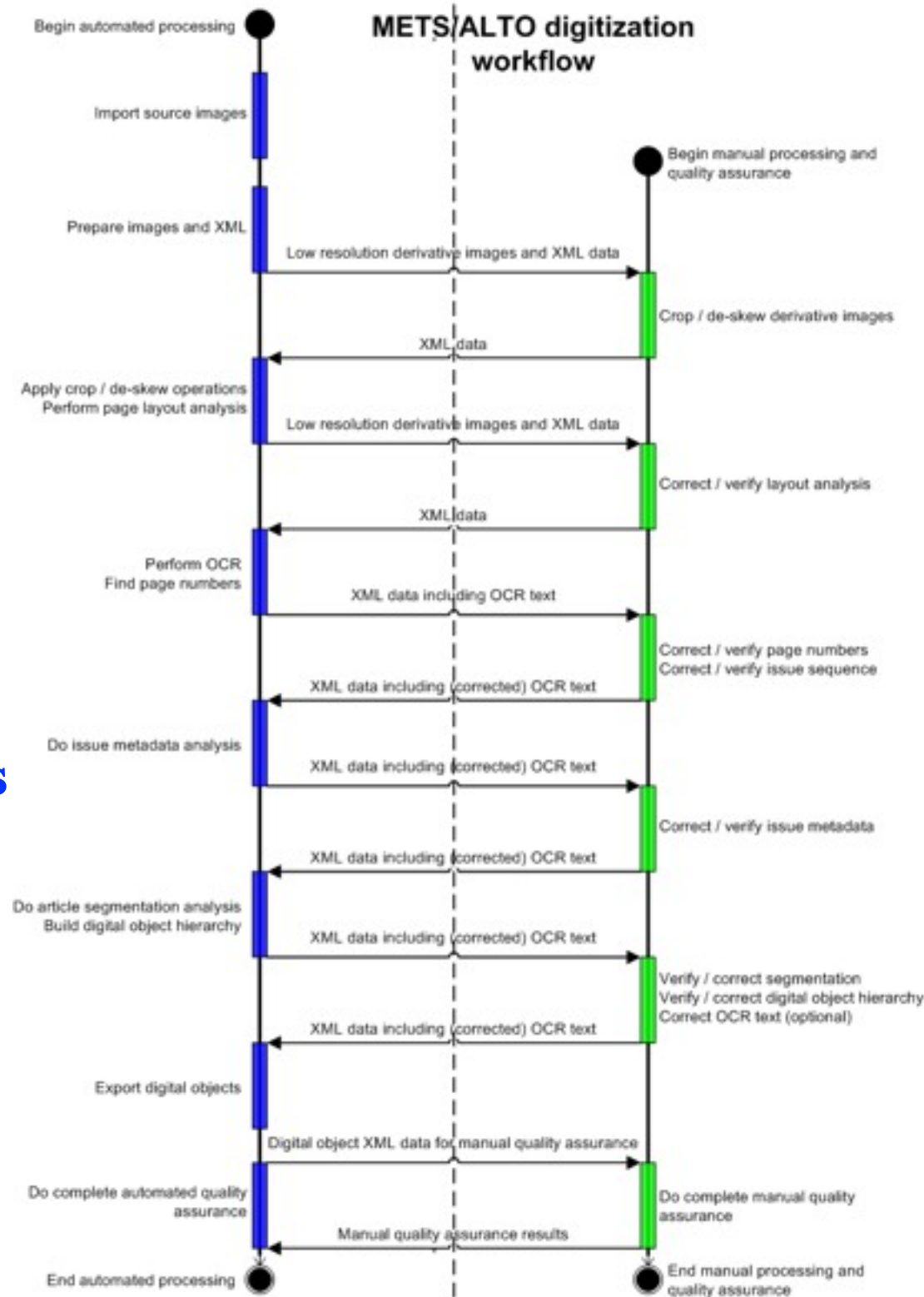
digitization magic



- create METS / ALTO XML files
- create image files and image metadata
- create PDF files (if required)
- verify digital object
- calculate file fixity checks (checksums)
- perform file validation and verification
- perform quality assurance

real world digitization production workflow

- automatic production steps performed by software
- manual production steps performed by operators



newspaper digitization programs around the world



National Library of Finland (<http://digi.kansalliskirjasto.fi/>)



British Newspaper Archives, British Library (<http://www.bl.uk/welcome/newspapers>)



National Digital Newspaper Program, Library of Congress
(<http://chroniclingamerica.loc.gov/>)



National Library of New Zealand (<http://paperspast.natlib.govt.nz/>)



National Library of Australia, Australian Digital Newspapers Program
(<http://trove.nla.gov.au/newspaper>)



Koninklijke Bibliotheek, the Netherlands (<http://kranten.kb.nl/>)



Singapore National Library Board (<http://newspapers.nl.sg/>)



Bibliothèque nationale de France (<http://gallica.bnf.fr/>)



Europeana Newspapers Project, a collaboration of 17 organizations (<http://www.europeana-newspapers.eu/>)

image references and recommendations

- Ian Bogus et al. *Minimum Digitization Capture Recommendations* (draft). The Association for Library Collections and Technical Services. June 2012 (accessed 18 Aug, 2012 at <http://connect.ala.org/node/185648>).
- Robert Buckley and Simon Tanner. JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library. Xerox Corporation and King's College Digital Consultancy for the Wellcome Trust Library. August 2009 (accessed 1 July 2012 at <http://library.wellcome.ac.uk/assets/wtx056572.pdf>).
- Paolo Buonora and Franco Liberati. *A Format for Digital Preservation of Images: A Study on JPEG 2000 File Robustness*. D-Lib Magazine. July/August 2008. (accessed 1 July 2012 at <http://www.dlib.org/dlib/july08/buonora/o7buonora.html>).
- ANSI/NISO Z39.87-2006. Data Dictionary -- Technical Metadata for Digital Still Images. National Information Standards Organization, Bethesda, Maryland USA. December 2006. (accessed 1 August 2012 at http://www.niso.org/apps/group_public/download.php/6502/Data%20Dictionary%20-%20Technical%20Metadata%20for%20Digital%20Still%20Images.pdf).
- JBIG Standard (accessed 1 August 2012 at <http://www.jpeg.org/jbig>).
- JPEG Standard (accessed 1 August 2012 at <http://www.jpeg.org/jpeg>).
- JPEG2000 Standard (accessed 1 August 2012 at <http://www.jpeg.org/jpeg2000/>).
- TIFF 6.0 Standard (accessed 1 August 2012 at <http://partners.adobe.com/public/developer/tiff>).
- Many, many others....

newspaper digitisation references



Australian Newspapers Digitisation Program

<https://www.nla.gov.au/ndp/>



europeana
newspapers

Europeana Newspapers

<http://www.europeana-newspapers.eu/>



IFLA Newspapers Section

<http://www.ifla.org/en/newspapers>



IMPACT Centre of Competence

<http://www.digitisation.eu/>



Koninklijke Bibliotheek
Historische Kranten

Koninklijke Bibliotheek Historische Kranten (the Netherlands) <http://kranten.kb.nl/about>



Library of Congress National Digital Newspaper Program

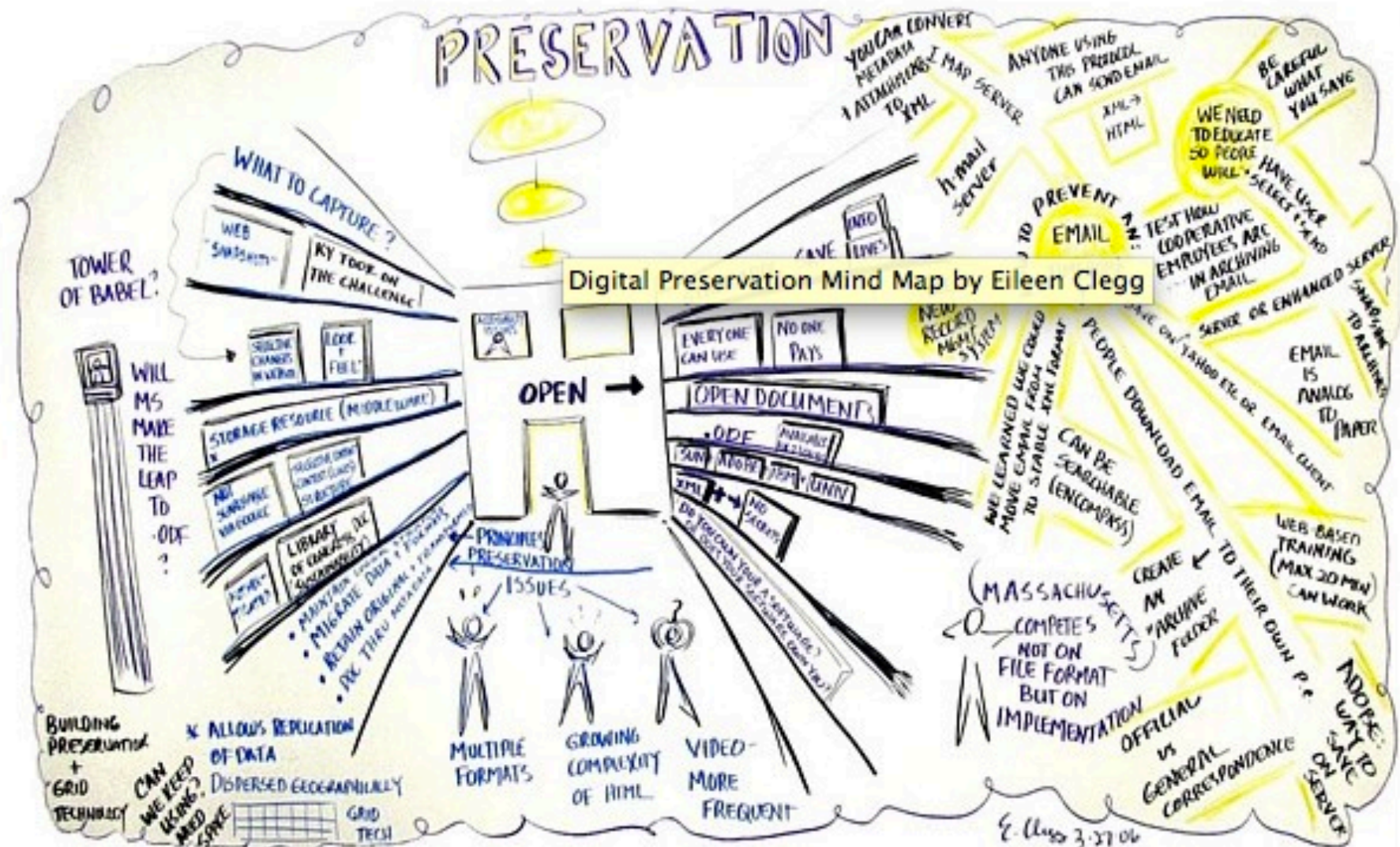
<http://www.loc.gov/ndnp/>



Frederick Zarndt
Chair, IFLA Newspapers Section
frederick@frederickzarndt.com

Part 2
***Brief words about digital
preservation***

digital preservation



digitization

digitization

***digital
preservation***

digitization

≠

*digital
preservation*

digitization

≠

*digital
preservation !*

digital preservation

long-term, error-free storage of digital information, with means for retrieval and interpretation, for the entire time span the information is required

digital data risks



- bit rot
- media obsolescence / decay
- migration to new format, media, or hardware
- standards / format obsolescence



bit rot

gradual decay of ...

- storage media because of media quality
- storage media because of improper storage
- data due to random events (bit-flip, environmental influences)

prevention / detection of bit rot

- data file checksums (MD5, SHA-1, ...)
- monitor file integrity (re-compute file checksums and compare)
- duplicate copies, geographically distributed (LOCKSS, CLOCKS, Chronopolis, ...)

media decay

a report by NIST and the Library of Congress says ...

- virtually all CD-Rs tested indicated an estimated life expectancy beyond 15 years
- only 47 percent of recordable DVDs indicated an estimated life expectancy beyond 15 years, some had a life expectancy as short as 1.9 years
- in practice actual lifetimes may be considerably shorter

prevention / detection of media decay

- proper storage
- data file checksums (MD5, SHA-1, ...)
- monitor file integrity (re-compute file checksums and compare)
- migrate data from old media to new media

media obsolescence

- 5 1/4" floppy disks
- 8 track tapes
- 3 1/2" floppy disks
- ZIP drives
- CD-R, CD-RW, Blu-Ray
- DAT tapes
- microfilm
- etc

strategies for media obsolescence

- migrate data to new media, for example, floppy disks to DVD
- create and maintain a computer hardware museum

data migration risks

- file format changes, for example, PDF 1.4 to PDF 1.8
- file name differences, for example, case sensitive /insensitive names, new operating system
- extended file attributes
- file permissions, for example, BSD Unix *drwxr-xr-x@* to Windows
- soft links / hard links

format obsolescence

remember ...

- WordPerfect ?
- MARC records ?
- Adobe Flash ?

strategies for format obsolescence

- migrate data to new formats
- create a computer software museum (virtual machines)

distributed decentralized preservation

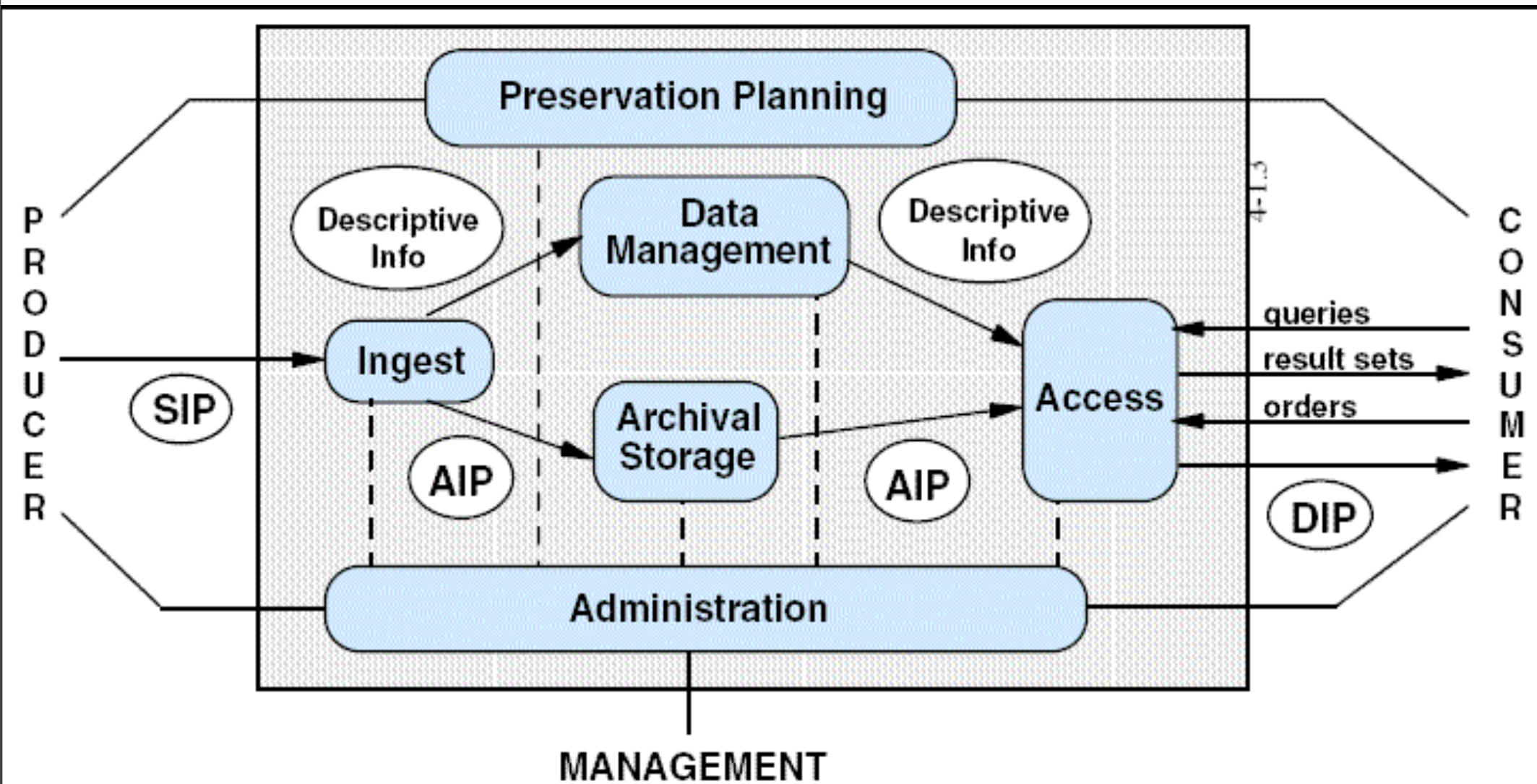


LOCKSS

Lots Of Copies Keeps Stuff Safe

preservation

Open Archival Information System (OAIS) reference model





Frederick Zarndt
Chair, IFLA Newspapers Section
frederick@frederickzarndt.com

Part 3
***The importance of
communication,
specifications,
acceptance criteria***

the problem

Wise men learn by other men's mistakes, fools by their own.
H. G. Wells

the problem

the 2009 CHAOS Report (The Standish Group) reports that of all software projects surveyed, 44% are “challenged”, 24% failed, and only 32% succeeded

the problem

Roger Sessions estimates that the worldwide cost of IT failure is USD \$500 billion per month

Roger Sessions: CTO of ObjectWatch. He has written seven books including *Simple Architectures for Complex Enterprises* and many articles. He is a founding member of the Board of Directors of the International Association of Software Architects.

the problem

in a recent survey of 1230 IT professionals conducted by Embarcadero Technologies, 2 of the 3 biggest project challenges cited by the IT pros are “poor planning” and “poor or no requirements”

the problem

in a March 2007 web poll conducted by the Computing Technology Industry Association "nearly 28 percent of the more than 1,000 respondents singled out poor communications as the number one cause of project failure"

the problem

in a white paper written for Project Perfect by Taimour al Neimat, he lists

- *poor planning*
- *unclear goals and objectives*
- *objectives changing during the project*
- *unrealistic time or resource estimates*
- *lack of executive support and user involvement*
- *failure to communicate and act as a team*
- *inappropriate skills*

as primary causes for the failure of complex IT projects

the problem

a recent tender from an (anonymous) government agency

- *project to convert ~ 170,000 text images to xml*
- *value of project ~ USD \$180,000*
- *19 pages of definitions, governing law, proposal evaluation criteria, contractual conditions, instructions about tender response format, etc*
- *technical requirements description? < 1 page*
- *data acceptance criteria? “a high level of accuracy”*

the problem

a recent program established by a prominent national library

- *digitize more than 20 million text pages*
- *high level image and xml requirements*
- *value of work awarded? > USD \$5,000,000*
- *after award of work, METS xml technical requirements expand to 43+ pages from ~3 pages*
- *acceptance criteria? added as an afterthought and not well defined*

the problem

acceptance criteria for a digitization program at a prominent library

character accuracy > 80%

word accuracy > 75%

significant word accuracy > 65%

the problem

typical tender evaluation criteria in priority order

- 1. understanding of requirements***
- 2. reputation of service bureau*
- 3. price*



the problem

communication

acceptance

requirements

the illusion

In theory, there's no difference between theory and practice, but in practice, there is.

Anonymous

The single biggest problem in communication is the illusion it has taken place.

George Bernard Shaw

the illusion waterfall requirements

for each product release repeat

{

gather requirements

create architecture

design

implement

test

use -or- sell

}

until (company goes out of business)

the illusion requirements

a recent tender from an (anonymous) government agency

- *project to convert ~ 170,000 text images to xml*
- *value of project ~ USD \$180,000*
- *19 pages of definitions, governing law, proposal evaluation criteria, contractual conditions, instructions about tender response format, etc*
- *technical requirements description? < 1 page*
- *data acceptance criteria? “a high level of accuracy”*

the illusion acceptance criteria

acceptance criteria for a digitization program at a large,
well-known, and internationally recognized national library

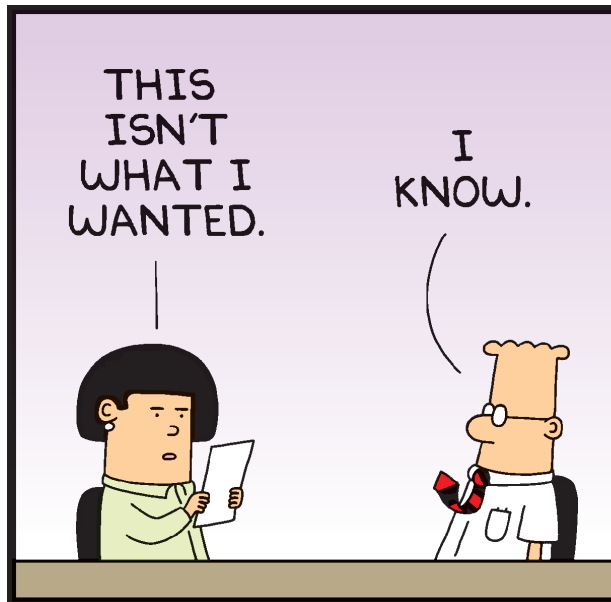
character accuracy > 80%

word accuracy > 75%

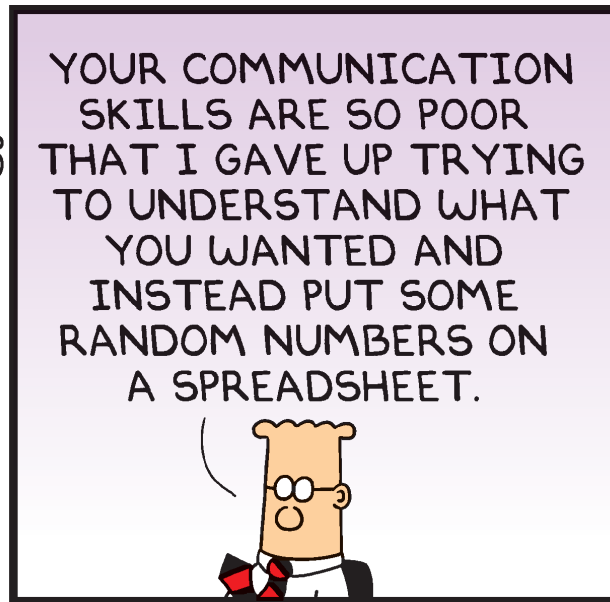
significant word accuracy > 65%

the illusion

why (better) communication is necessary



Dilbert.com DilbertCartoonist@gmail.com



4-2-11 ©2011 Scott Adams, Inc./Dist. by UFS, Inc.



the fix

Experience is that marvelous thing that enables you to recognize a mistake when you make it again.

F. P. Jones

the fix value of simplicity

*“Perfection is attained, not when there is nothing left to add, but when there is nothing left to take away.”
Antoine de St. Exupery*

the fix

value of prototypes and pilot batches

“Plan to throw one away; you will anyhow. If there is anything new about the function of a system, the first implementation will have to be redone completely to achieve a satisfactory (i.e., acceptably small, fast, and maintainable) result. It costs a lot less if you plan to have a prototype.”

Butler Lampson

Butler Lampson was a founding member of Xerox PARC, worked for DEC, and now works at Microsoft Research. He is an adjunct professor at MIT and an ACM Fellow.

the fix value of simplicity

“There are two ways of constructing a software design: one way is to make it so simple that there are obviously no deficiencies and the other way is to make it so complicated that there are no obvious deficiencies.”

C.A.R. Hoare

Professor Sir Charles Anthony Richard Hoare Emeritus Professor at Oxford University, Senior Researcher at Microsoft Research, recipient of the ACM Turing Award, author of many books on computers and software.

the fix

good requirements

- *unitary: the requirement addresses one and only one thing*
- *complete: the requirement is fully stated in one place with no missing information*
- *consistent: the requirement does not contradict any other requirement and is fully consistent with all authoritative external documentation*
- *atomic: it does not contain conjunctions, for example, "the code field must validate American and Canadian postal codes" should be written as two separate requirements*
- *traceable: the requirement meets all or part of a business need as stated by stakeholders and authoritatively documented*

the fix

good requirements (continued)

- *current: the requirement has not been made obsolete by the passage of time*
- *feasible: the requirement can be implemented within the constraints of the project*
- *unambiguous: the requirement is concisely stated without recourse to technical jargon, acronyms*
- *verifiable: the implementation of the requirement can be determined through one of four possible methods: inspection, demonstration, test, or analysis*

the fix *requirements and acceptance criteria*

Wikipedia on data quality: The processes and technologies involved in ensuring the conformance of data values to requirements and acceptance criteria

the fix
requirements and acceptance criteria

“a high level of accuracy”

the fix
requirements and acceptance criteria

“article titles must be 99.5% accurate”

the fix

requirements and acceptance criteria

“article title characters in each issue must be 99.5% accurate, that is, each issue may have no more than 5 errors in 1000 article title characters”

the illusion waterfall requirements

for each product release repeat
{
 gather requirements
 create architecture
 design
 implement
 test
 use -or- sell
}
until (company goes out of business)

the fix *agile requirements*

gather general requirements

create architecture

build prototype software

test

repeat

{

use software

adjust prototype and/or add new feature

test

}

until (user says stop or runs out of money)

the fix

agile data conversion

create requirements and acceptance criteria

repeat

{

digitize (small) pilot batch

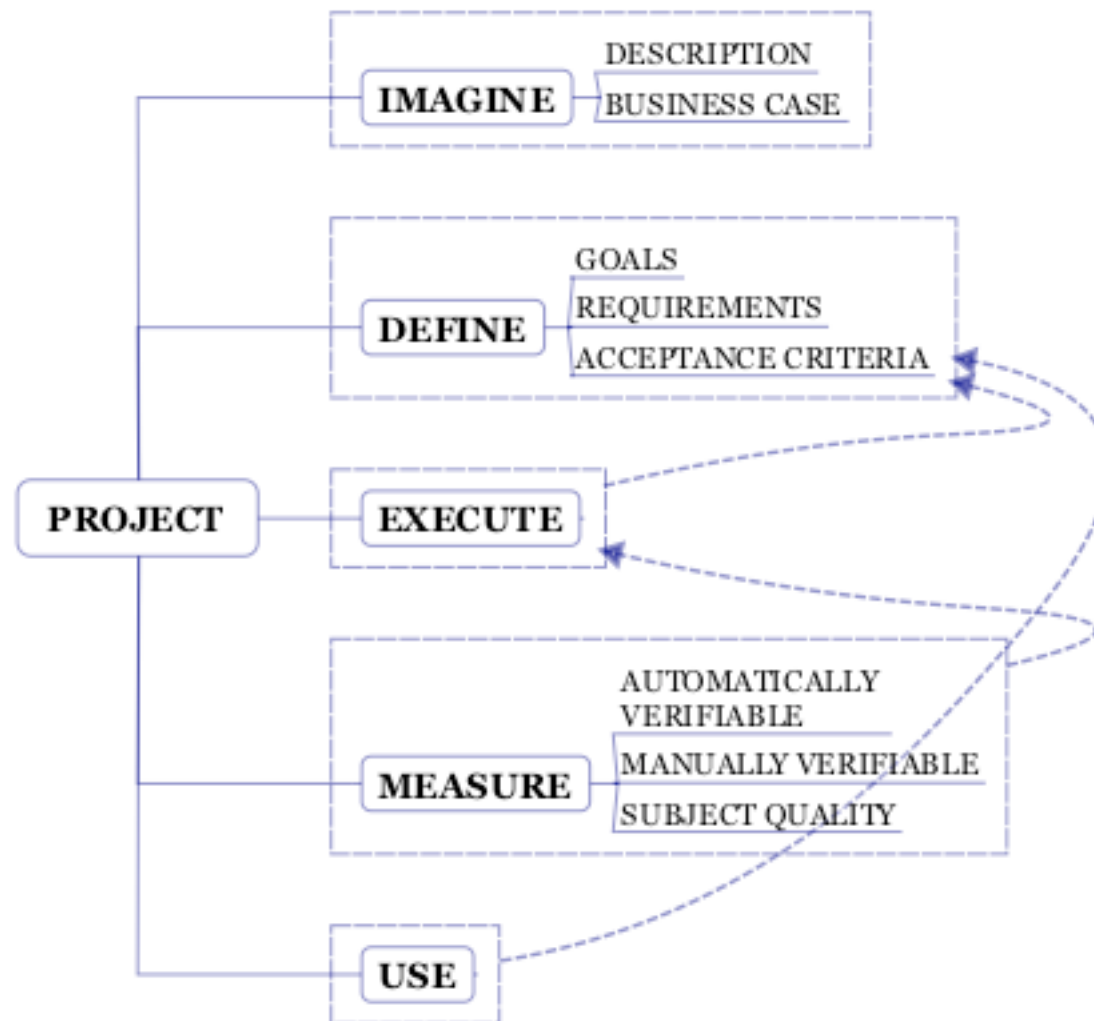
test data against acceptance criteria

adjust requirements and acceptance criteria

}

until (no more adjustments are necessary)

digitize more data



the fix
why (better) communication is necessary

*“projects are about communication,
communication, and communication”*

the fix

simple principles for (good) communication

- *be impeccable with your word*
- *don't take anything personally*
- ***don't make assumptions***
- *always do your best*
- *be mindful*

the fix
why (better) communication is necessary

no communication ...

the fix

why (better) communication is necessary

no communication ...

little communication ...

the fix

why (better) communication is necessary

no communication ...

little communication ...

poor communication ...

the fix

why (better) communication is necessary

no communication ...

little communication ...

poor communication ...

reduced communication ...

the fix

why (better) communication is necessary

no communication ...

little communication ...

poor communication ...

reduced communication ...

*... all result in more assumptions about
intent!*

the fix

how do you communicate?

- communication is **at most** 30% verbal!
- remainder - 70% or more - is comprised of gestures, facial expressions, tone of voice, posture, odors, ...
- telephone communication removes gestures, facial expressions, posture, odors, etc. only words and tone of voice remain
- written communication - email, requirements, etc - removes all modes of communication save for words

the fix

how to communicate

simple *keep it simple stupid (KISS principle)*

repeat *say it twice in different ways*

listen *repeat what you hear*

respect *respect yourself and others*

conclusion

for future projects give especial attention to

*good, open communication
clear requirements
clear acceptance criteria*



We all admire the wisdom of people who come to us for advice.
Jack Herbert

Frederick Zarndt
Chair, IFLA Newspapers Section
frederick@frederickzarndt.com