the well-behaved document

Presented by
John W. Miescher
Bizgraphic – Geneva
On Thursday, October 06, 2011

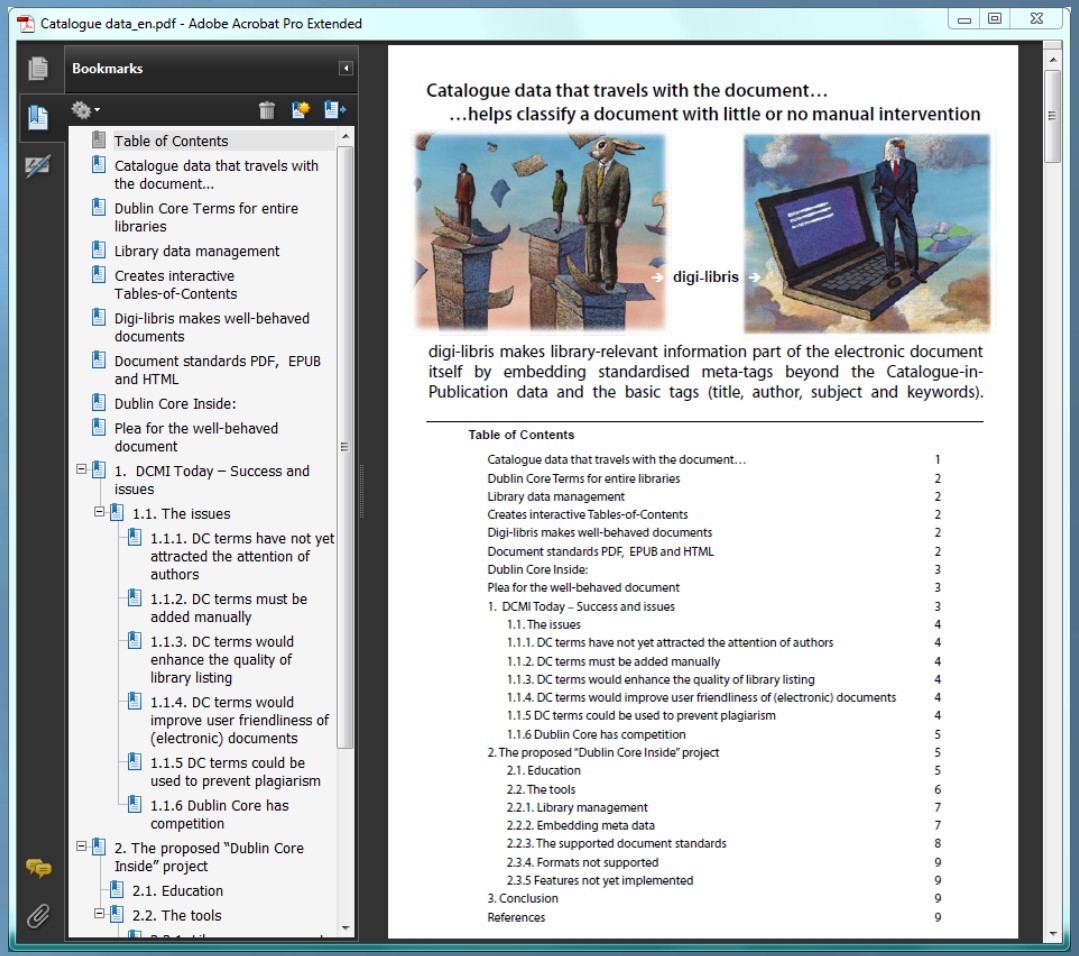
the well-behaved document

is an electronic document that is both user friendly **and** library friendly

is easy to read
and to navigate

it has bookmarks

and an interactive
table-of-contents



is practical to consult and arouses more interest

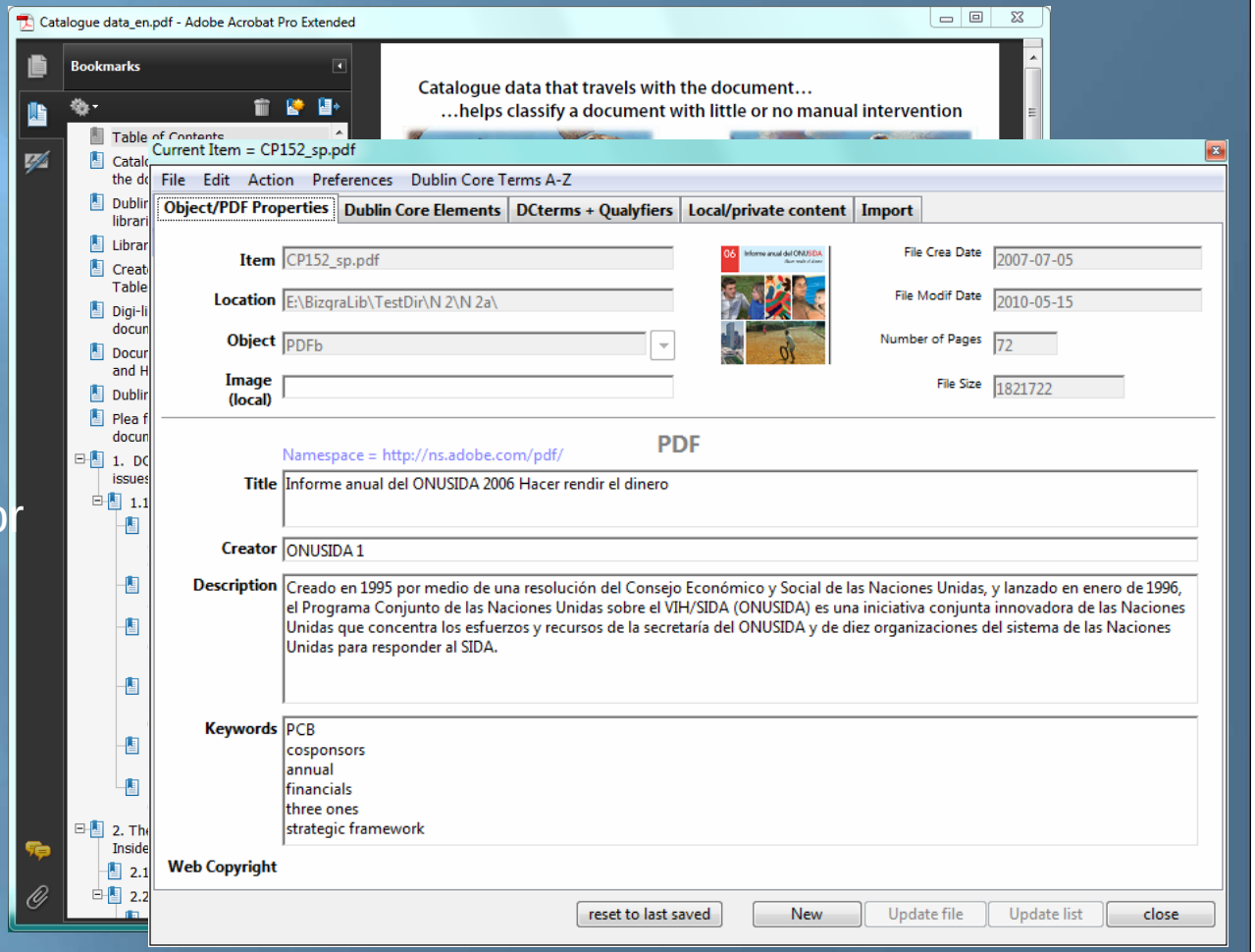
the well-behaved document

is an electronic document that is both user friendly **and** library friendly

it includes useful metadata as part of the document itself

to help classify it with little or no manual intervention

for personal data mining and re-use



are more attractive for readers to keep and refer to

have a better chance of being found (in search engines)

Who needs it?

author, researcher

faculty, publisher, self, librarian, archivist, digitizer, asset manager, etc.

you, me, students, scholars, non-scholars (min. 75%)

about Open Access

Open Access is migrating from a buzzword to a quasi standard for scientific publishing.

While emerging concepts like Nanopublikationen, Semantic Publishing, Open Data, Enhanced Publications and Research Objects all have merits in scientific context, Google Scholar might end up stealing the show, even in research.

Great efforts are being made to streamline ontologies (SPAR, CiTO), semantics (CERIF, CRIS), protocols and data integration (SWORD, SHERPA/RoMEO), university/research libraries grouping together to standardize and coordinate their repositories and opening data licensing (LIBER, CENL, CERL, OAPEN, COAR, openAIRE, openDOAR etc.) and

metadata harvesting and interoperability becoming an increasingly important issue (OAI-PMH, MARC, Dublin Core).

All this is undertaken by university and library professionals for the benefit of university and library professionals, it seems.

While the benefits to the individual information consumer appear to be obvious, we were not able to find many pieces of instruction targeted at the end user on how to best profit from these efforts and what skills he/she needs to develop in order to maximize his/her search for information.

why does he/she need it?

looks-up data or collects information

author, researcher, scholar or occasional writer.

searches the Web
finds hundreds of references

spends time collecting, sorting and referencing citations and data from mostly unstructured text

must decide on relevant and essential information

and would find embedded metadata most helpful

if you are doing research work

if all documents were “well-behaved”…

had bookmarks and interactive tables of contents
and had meaningful embedded metadata

the consumer would have easy access to information

to refer to and to consult off-line sources
He/she has acquired or downloaded

and the writer of a thesis, a report or a paper

wouldn’t waste time in search of relevant information,
collecting citations and assembling data

about embedded metadata

The average information consumer…

is primarily interested in the descriptive metadata and less in the structured and administrative metadata

does not care about semantics, namespaces and refinements

the Dublin Core terms (DCMI interoperable online metadata) is probably the best option for embedding useful metadata

the format is widely recognized and well documented

can easily be embedded in PDF, EPUB and HTML files

in PDF files it is already included as standard

more about Dublin Core

DC terms have not yet attracted the attention of authors

authors are non-librarians

authoring tools allow at best only basic metadata

the discipline to fill in even these basics is not there

15 CD-elements and 55 DC-terms can lead to confusion

Dublin Core is not alone

there are lots of standards issued by major libraries, universities, school authorities and other interest groups

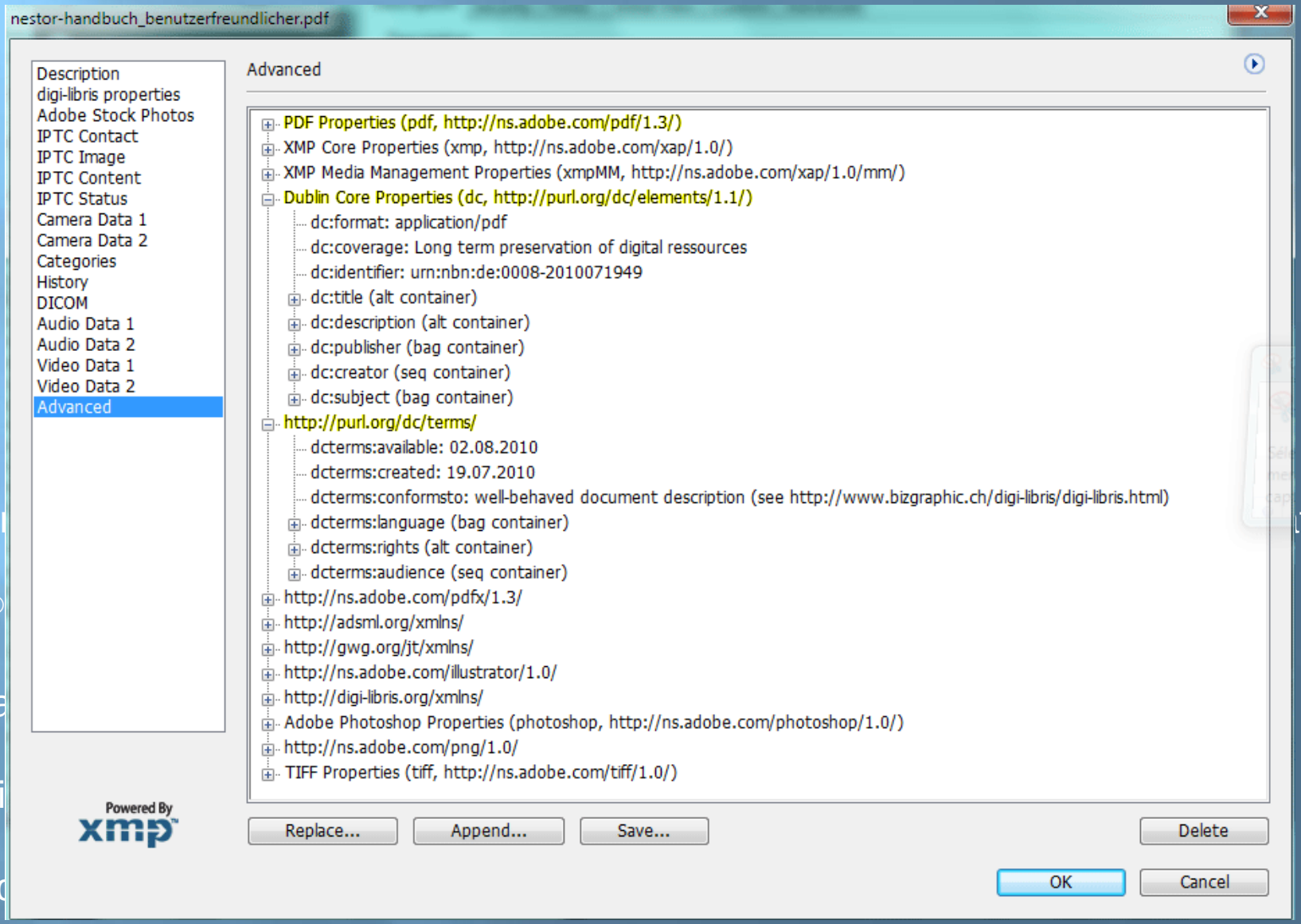
MARC 21 is used by major libraries for their own repositories

Unfortunately it is not very compatible with the Dublin Core standard

and the *MARC to Dublin Core Crosswalk* has its limits

embedding Dublin Core tags

PDF



most scan

Acrobat®

has a spa

includes i

but no too

embedding Dublin Core tags

EPUB
the new standard for ebooks.
is not proprietary
some layout programs (e.g. InDesign) can generate it directly
Handled by most major ebook readers
Most examples we have seen use dc:terms correctly.

HTML
the most uncontrolled format of all
there are hundreds of applications that can create it, *each with his own flavor of syntax and scripts*

By far not all originators use <dc...> or <dc.terms…> for metadata
and often those that do add fancy designations of their own.

Can be virtual or created on the fly, *e.g. in response to a query*

helpful software tools

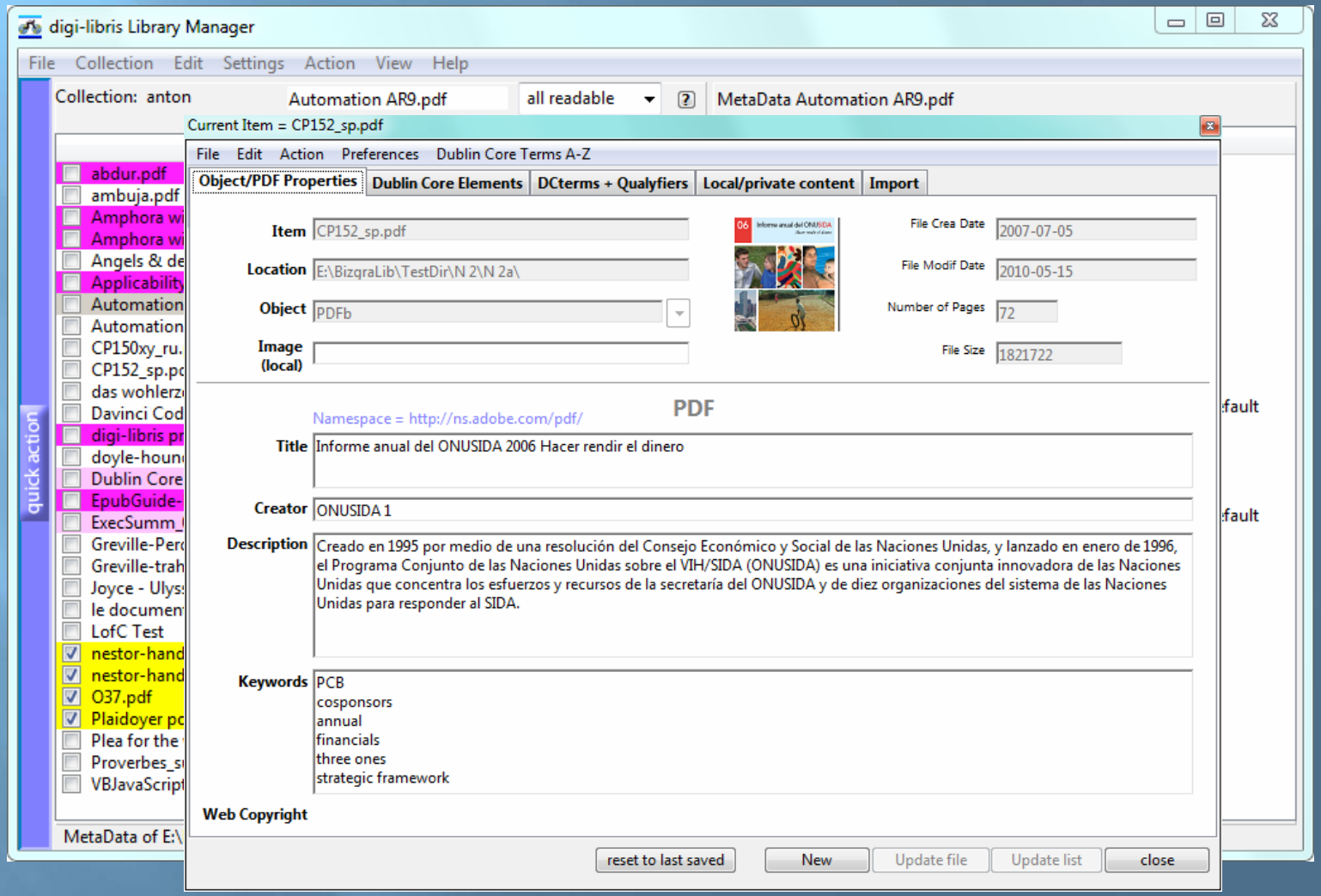
digi-libris to organize your knowledge base and collections of documents

add physical or electronic items to your collection

and see the embedded metadata

or edit and re-embed the metadata

to make a document library friendly



download a trial version of this software from http://digi-libris.com

post production software

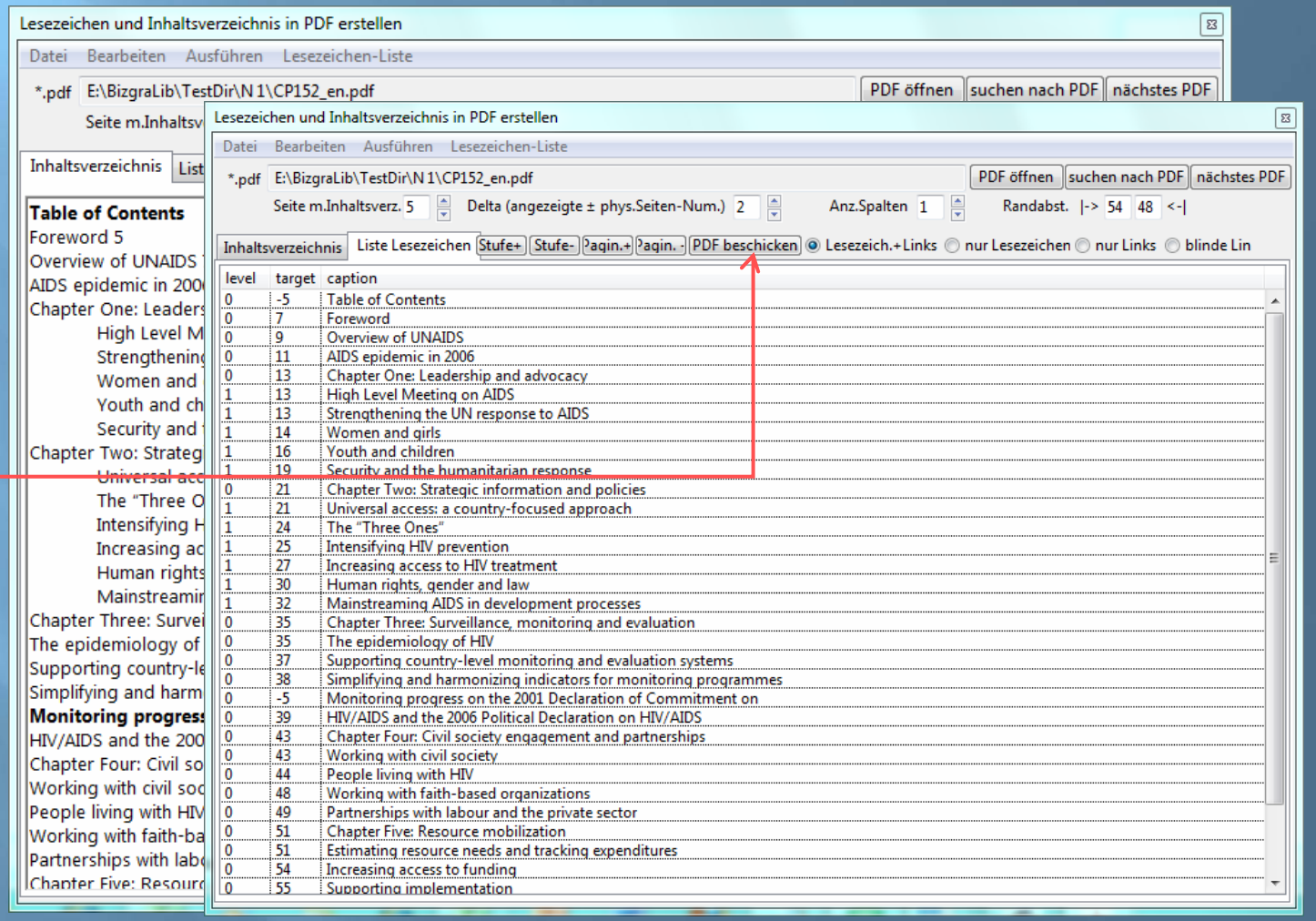
to enhance existing PDF files, e.g. third party or digitized ones you can…

copy/paste or type a table of contents

then automatically generate and edit a bookmark list

and send * bookmarks and links to a PDF to to build an interactive table of contents

to make a document user friendly

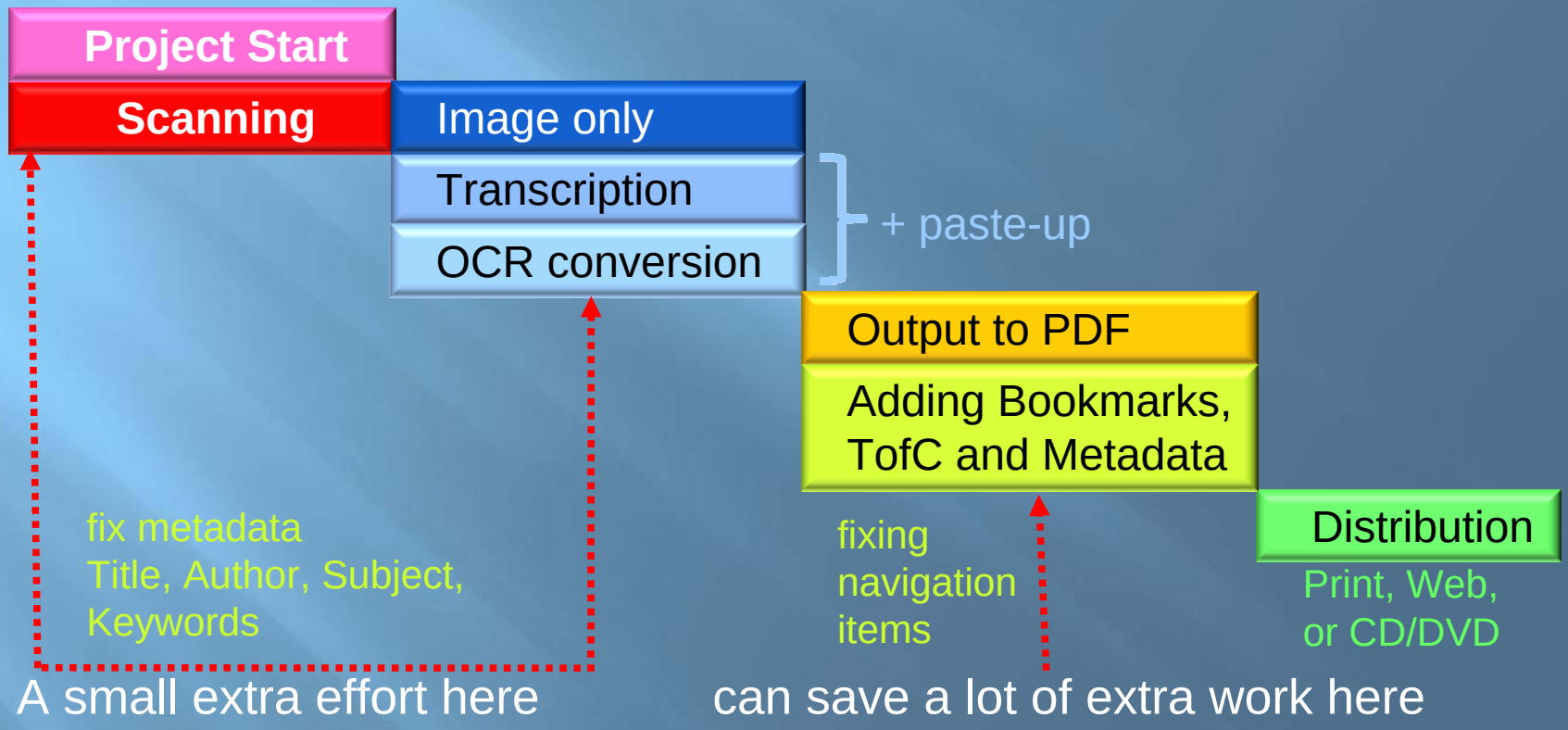


download a trial version of this software from http://digi-libris.com

* requires Acrobat©

good planning saves time and money

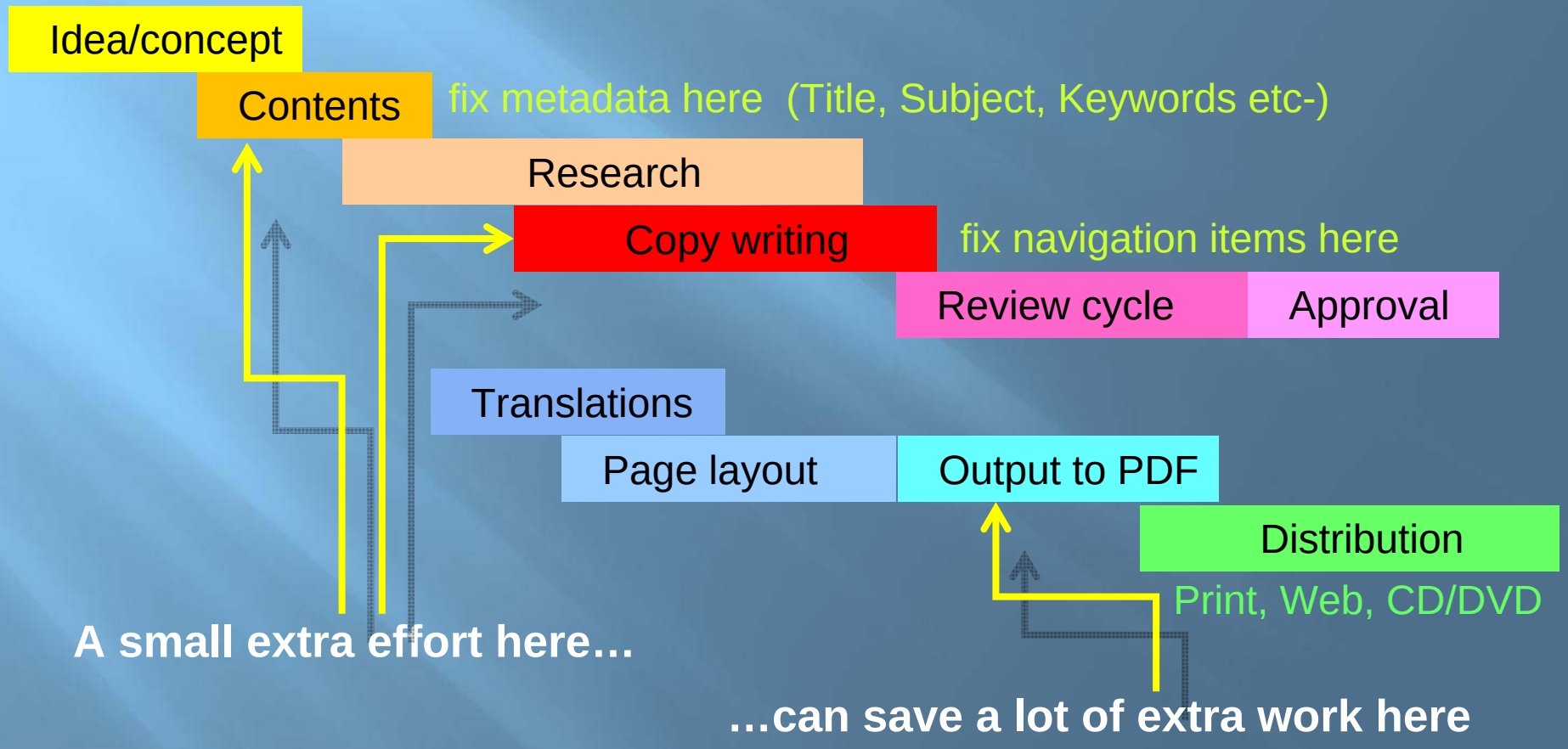
Consider the workflow of a typical digitizing project



to make your document a well-behaved one

good planning saves time and money

Consider the workflow of a typical document creation process



To make your document a well-behaved one

Well-behaved documents…

cater to the needs of (and empower) the information consumer

have a better chance of being found (in search engines)

facilitate personal data mining and re-use

are practical to consult and arouse more interest

are more attractive for readers to keep and refer to repeatedly

are a step in the direction of the semantic web

allow processing of library records with little/no manual intervention

can save time and money if planned and implemented accordingly

the well-behaved document has Dublin Core inside

Introducing “Dublin Core inside” as a new standard and mark of quality for truly well-behaved documents,

to be recognized worldwide by information providers and information consumers alike.

How?

by using our influence, relations and know-how with authors, publishers, aggregators and content providers

by helping all stake holders to understand and appreciate the added value of well-behave documents and to prepare all their documents accordingly.