Gigabytes, Terabytes, Petabytes........Now What?

An Automated Approach and New Tools for Managing Risks That Threaten Permanent Access

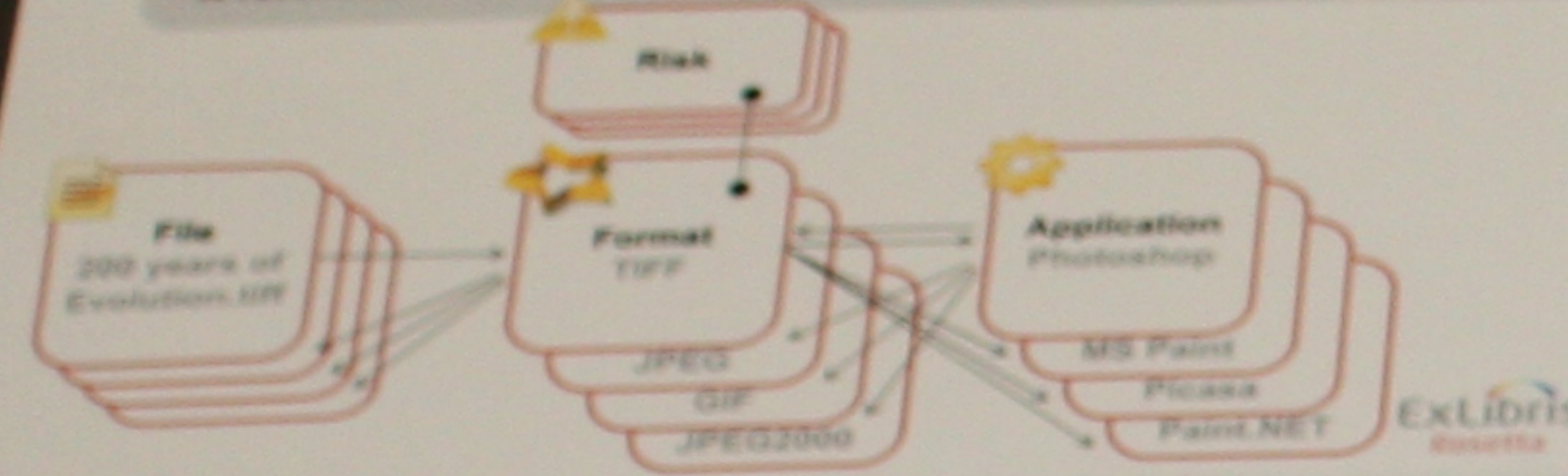*Mike Thuman, Ex Libris Director, Digital Preservation*

Current State

Is there a risk that your files will not be viewable?
- Institutions typically have tens of thousands to hundreds of millions of files (based on many different file formats)
- Hundreds of file formats to render (550+ in PRONOM Registry)
- An application can open and display one or more formats
- A format can have more than one risk

Risk Types

Risk
- Obsolete Format
- Application not found – no license, application end of life by vendor
- Files attributes – number of colors / MIME type / compression level

Software upgrades fail to support legacy files.

The format itself is superseded by another or evolves in complexity.

The format "take up" is low or industry fails to create compatible software.

The format fails, stagnates, or is no longer compatible with the current environment.

Software supporting the format fails in the marketplace or is bought by a competitor and withdrawn.[1]

[1] http://www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/obsolescence1.html

---

Format Properties and Property Driven Risks

Format
TIFF
Properties:
- Mime Type
- exif.colourSpace
- image.pixelAspectRatio
- mix.compressionScheme

Different formats have different property types

Some property types are used by more than one format

Format
RTF
Properties:
- Mime Type
- exif.colourSpace
- text.unicodeBlocks
- text.lineEndings

Format
TIFF
Properties:
- Mime Type
- exif.colourSpace
- image.pixelAspectRatio
- mix.compressionScheme

Risk
Files with compression schema = 5 can't be opened

Files with specific property values are at risk

Risks are defined by a regular expression

Abstract

Institutions from around the world are reporting continued growth in collection sizes reaching not only into terabytes but now petabytes in size. Mitigating risks associated with technology changes like file format obsolescence can be a challenge for any size collection but it becomes even more challenging as our collection sizes increase and archival staff has limited time to develop needed tools and automation approaches.

Through an international collaboration with the National Library of New Zealand, Ex Libris is able to report on an innovation and approach that archives can take to manage and mitigate risks related to digital formats. New tools make it possible to create a test migration and then at the discretion of the archivist, create a full migration, moving at risk digital records to a safe and up to date state while maintaining all provenance.

A library of digital formats will enable professionals to manage at risk formats in a global registry and to register different risks for new formats providing an opportunity for the global community to share in their findings and share in the advancements while reducing risk and providing relief from a potentially time consuming, manual approach.

Method

- International collaboration
- Development Partner-National Library of New Zealand
- Build on research- e.g. Planets Initiative
- Develop a preservation planning module
- Look beyond just the migration action itself

Result-Preservation Planning Module

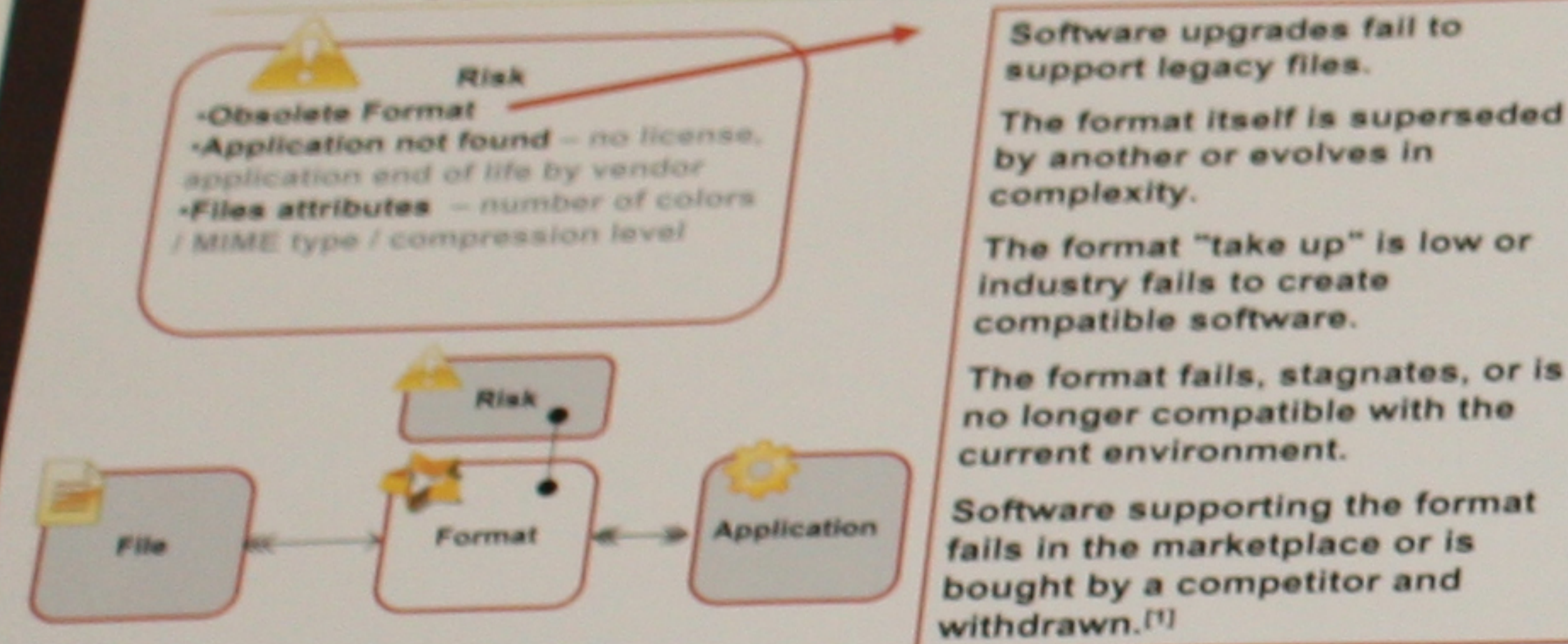Rosetta includes a preservation planning module that comprises
- a format library
- risk analysis
- evaluation and preservation actions

Allows institutions to manage and carry out the full preservation process of identifying risk, selecting the best among possible alternative solutions and testing/activating the preferred preservation migration process

---

Knowledge Base - Format Library

A library of different formats with metadata on each format

Based on the PRONOM global library

Allows adding and managing additional formats

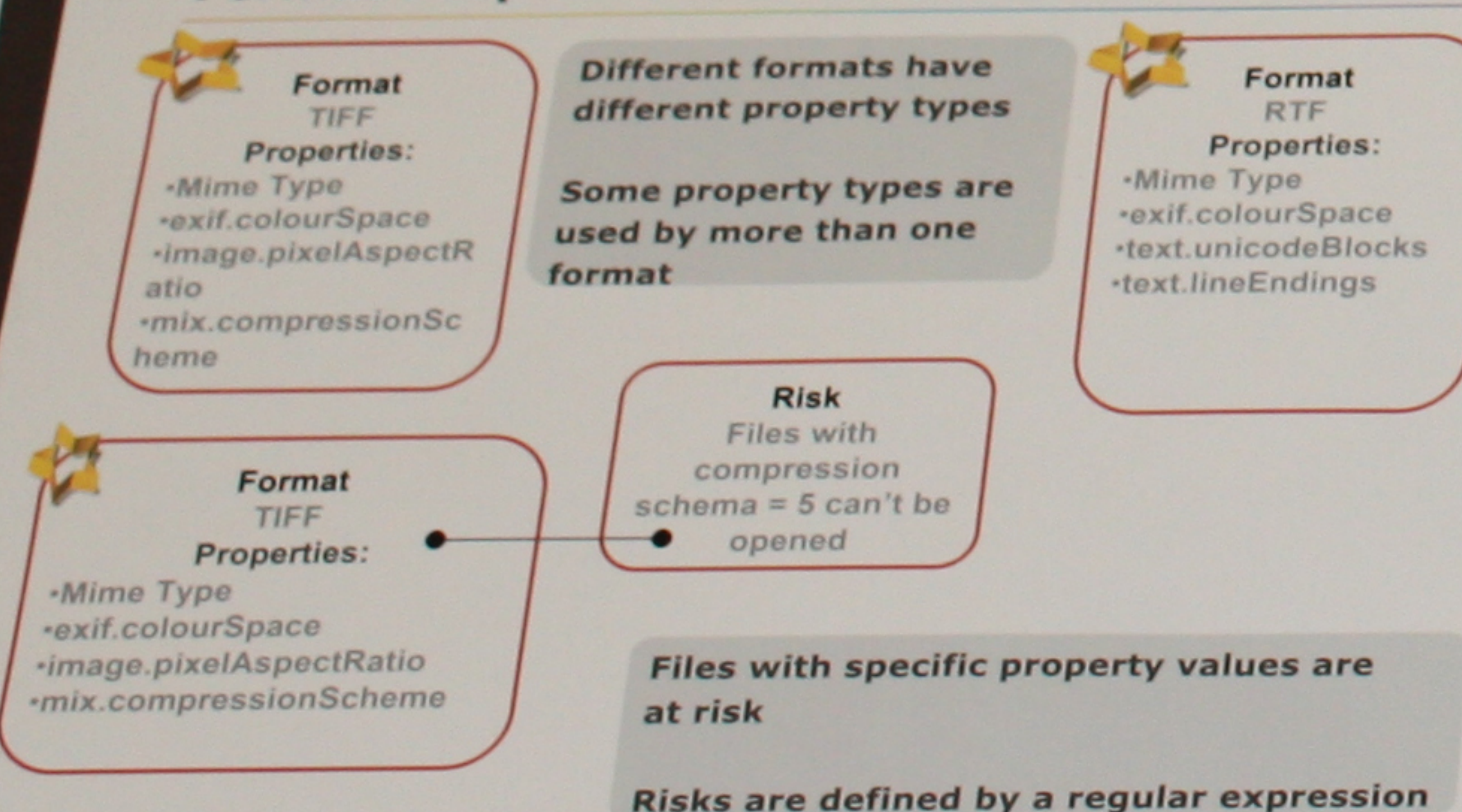Each format is associated to applications and risks

Supports integration with a global library

Global-of interest to all institutions using Rosetta (e.g. format becomes obsolete)

Local-of interest only to a specific institution (e.g. viewing application license has expired or format is not supported)

Global Library-an opportunity for rapid advancement and collaboration

---

Overall Summary/Results

Preservation planning doesn't start from the action itself.

Its a process which includes: format identification, metadata extraction, risk identification, evaluating alternatives, running the migration action and other steps

Preservation planning is largely about formats and risks that should be shared among institutions around the globe

Rosetta and the Global Library is an infrastructure that allows that to take place today

Digital preservation is still a growing issue. More research, collaboration, sharing and standardization are necessary for success