MAMK
University of Applied Sciences

Service oriented archive based on Fedora Commons

Mikko Lampi
Mikkeli University of Applied Sciences

MAMK
University of Applied Sciences

Preview

1. Background
2. Project drivers and design goals
3. Fedora and other building blocks
4. Development (and a few words on the ideology)
5. Future
6. Review and conclusion

MAMK
University of Applied Sciences

Background

Digital preservation and archiving in MAMK

- Fifteen years of research and development
  - Digital archive and repository software
  - Methods and tools development
  - Digitization, 3D scanning and modeling
  - Audiovisual materials
- Commercial digital archive services since 2004
  - Private archives and companies, city archives, non-profit organizations
  - Digital archive and repository as a service
  - Digitization, media productions
- Disec
  - Spin-off company for medical sector image archives and digital services
  - Provides MAMK an enterprise level infrastructure and data security

MAMK
University of Applied Sciences

Background

OSA - Open Source Archive

- Find and develop open source tools for digital preservation, repositories and archives
- Focus on developing a service platform for archives
- Pilot test a dark archive solution (DAITSS)
- Implementation during 2012 - 2014
- Funded by European Regional Development Fund, South Savo Regional Council, MAMK and partners
- Results will be released as much as possible open source
- Project blog: http://osarchive.wordpress.com/

We know how to do it.
This has been there for the last 5000 years.

MAMK
University of Applied Sciences

Motivation

- Upgrade current digital archive software
- Support changing requirements and agile development model
- Get rid of closed and proprietary software
  - Cut costs and understand the licensing better
  - Reduce risks and be in control
  - Political reasons (public sector, EU)
- A new architecture design
  - Modularity and loose coupling
  - Open source components
  - Flexible data models
- Provide top notch end-user experience

MAMK
University of Applied Sciences

Service model

- MAMK is a digital archive (and repository) service provider
  - SaaS (Software as a Service) with multi-tenant applications
  - Agile and user focused development
  - Focus on software and infrastructure, not in the content
  - Current production software is in-house developed YKSA
- Research and development projects integration
  - Continuum and funding outside of the projects
- Partnerships
  - such as ELKA (Central Archives for Finnish Business Records)
- Content agnostic services
  - Audiovisual materials
  - Documents
  - Maps, posters …
  - OAIS packages etc.

MAMK
University of Applied Sciences

Digital archive data lifecycle

Lots of processes in different phases of data lifecycle

- Ingest, migration, fixity, disposal etc.
- Some are organization dependant, some are not
- Configurability without added complexity – is it possible? Oh yes.

Lifecycle phases can be managed with workflows and plans

- Automation eliminates human errors and enforces processes
- Can be compared and shared with the community
- Micro-services based implementation

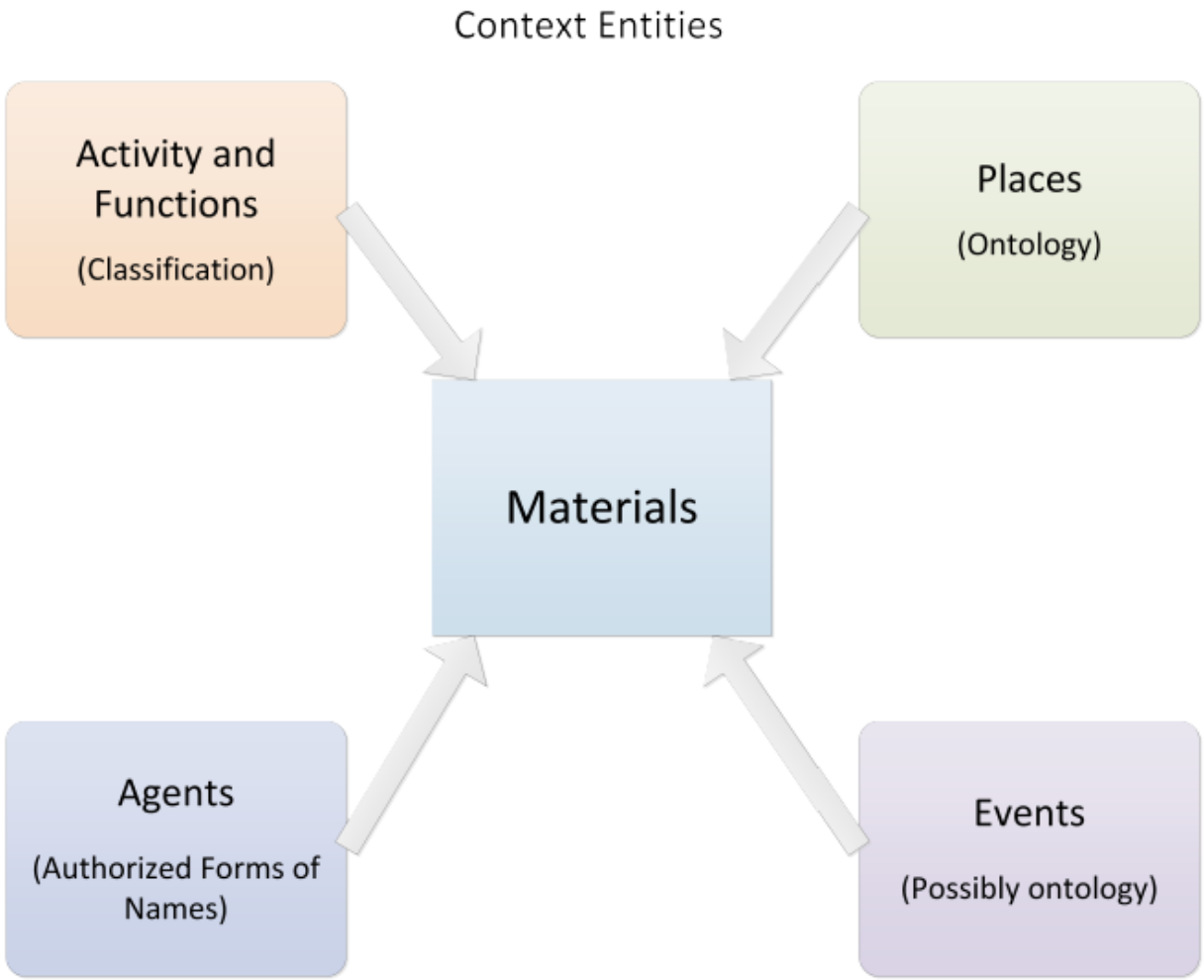
Digital archive or a repository is not a data tomb.

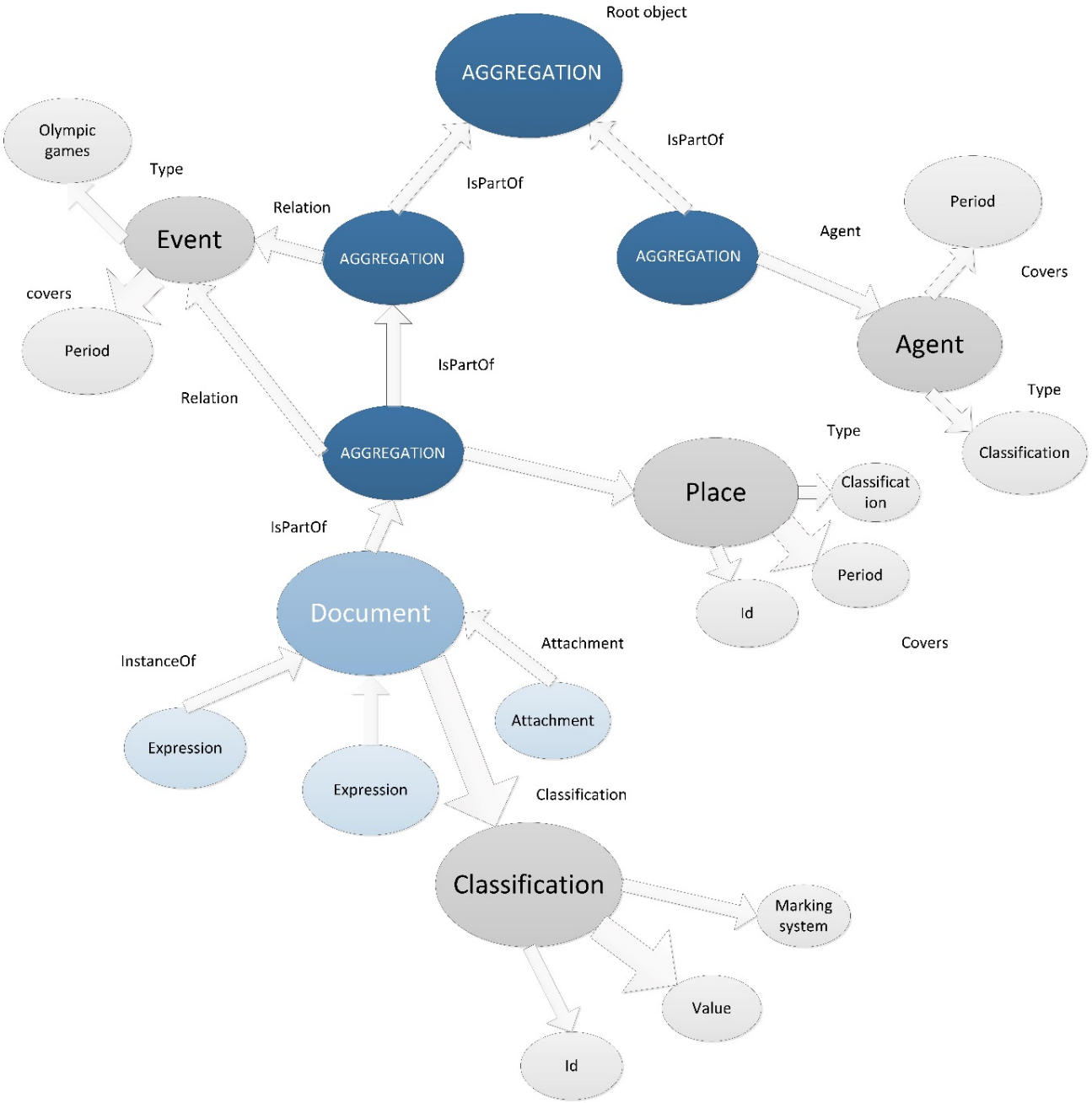
MAMK
University of Applied Sciences

Data modeling

- Very pragmatic approach
  - Archive first, enrich and enforce later
  - Do not limit the content or formats
- Umbrella metadata model
  - Covers multiple national and international standards
  - Roughly 300 metadata fields to cover various content types
  - Provide compatibility with mappings (which can be archived too)
  - Can be extended
- Machine readability
- Linked data
  - Internal and external (ontologies, classifications, vocabularies etc.)
  - Contextual entities

MAMK
University of Applied Sciences




Root object
AGGREGATION
Olympic games
Type
Event
Relation
AGGREGATION
IsPartOf
IsPartOf
AGGREGATION
Period
Agent
Covers
covers
Period
Agent
IsPartOf
Relation
Type
Classification
AGGREGATION
Type
Place
Classificat ion
IsPartOf
Period
Document
Id
Attachment
Covers
InstanceOf
Attachment
Expression
Expression
Classification
Classification
Marking system
Value
Id

MAMK
University of Applied Sciences

Discoverability and access

All data should be accessible and discoverable

- Without any knowledge of archive hierarchy etc.
- Natural language understanding
- Multilanguage support
- Google (like) searching
  - Downside is every results page after the first
- Faceted search and browsing based on metadata
- Linked data and open data
- Access control and privileges

MAMK
University of Applied Sciences

Research and evaluation

Done to avoid unnecessary re-inventing in 2012.

- Key requirements
  - Previous drivers
  - Open source
  - Active community and healthy ecosystem
  - Stable and reliable product
  - Good architecture and technical design
  - Flexible and customizable
- We ended up with Fedora and a few others (Hydra, Islandora, Archivematica).
- In the end techies decided. Fedora it was.

MAMK
University of Applied Sciences

Solution overview

- Fedora Commons as central repository
- Solr for search and indexing
- Custom developed front-end and business logic layer
- Java as core technology
  - Easy to find developers
  - Plenty of tools available
- MVC and service oriented architecture
  - Extendable and modular design
  - Loose coupling
- Disk and tape storage
- Runs on Linux

MAMK
University of Applied Sciences

Fedora Commons

- Currently Fedora 3.6.x
- Looking to start F4 testing during summer

- Why Fedora?
  - Technology base (such as Java, APIs)
  - Community and use cases
  - Object modeling
  - Content and data model agnostism
- Role of Fedora in our solution
  - Master data storage
  - Low-level storage management
  - Manages audit logs, versions, relations, compound objects
  - Basically keeps it all together

MAMK
University of Applied Sciences

Experiences with Fedora

What we did

- Created Custom content models
  - Looked for Islandora and other examples
  - Based on content types
  - Defined minimun requirements (metadata, relations, data streams)
  - Designed schemas for metadata models
- Interfaces (APIs or GUI) provide mappings per customer
- UI elements (forms, views) are completely configurable and decoupled from the content models

What we didn’t like, use or understand

- SOAP API
- Service definitions and deployments
- Hard coded policies e.g. access rights

MAMK
University of Applied Sciences

Open source compontents

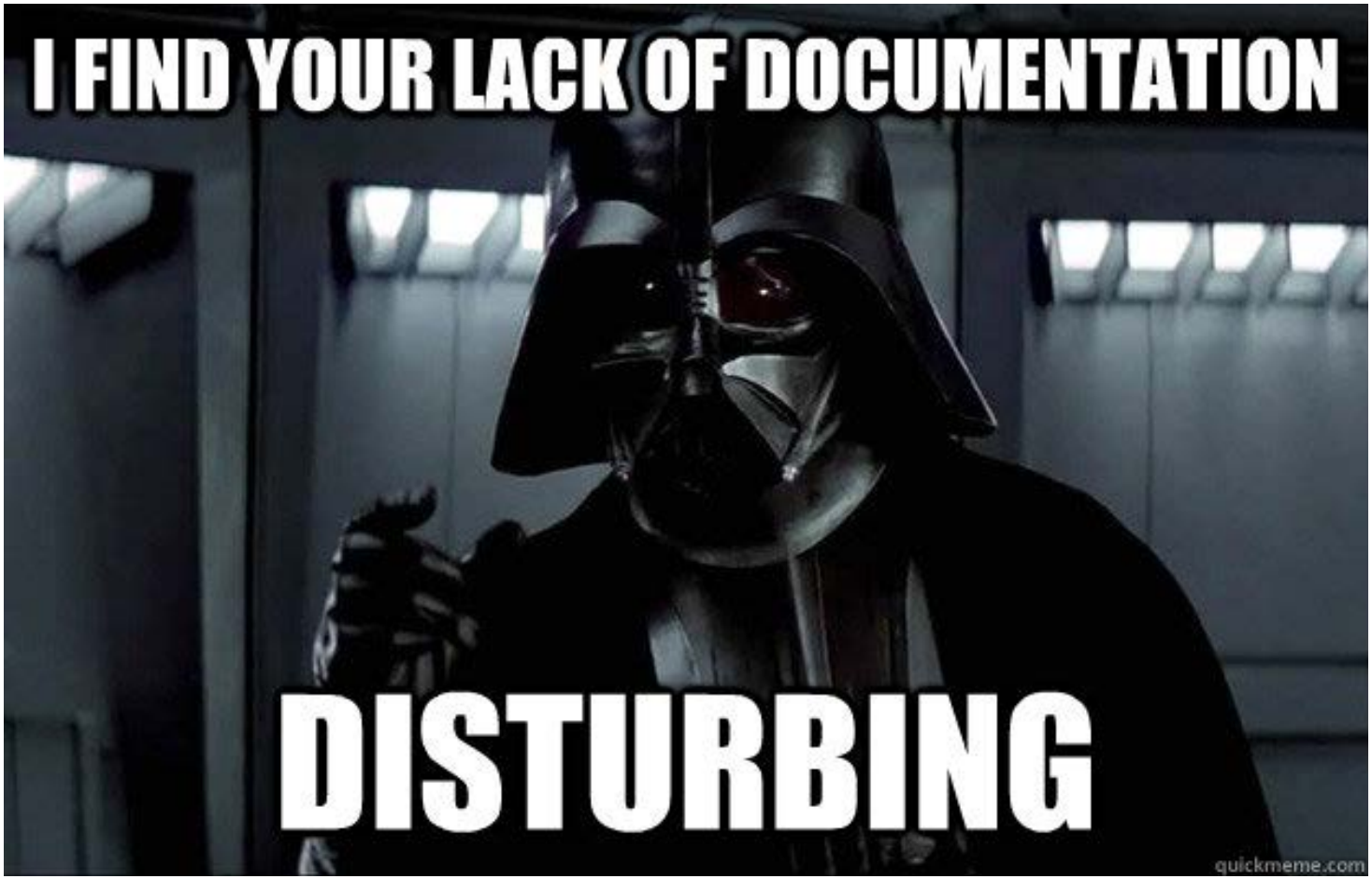
- Apache Solr 4.x
  - Gsearch (moving away with F4 adoption)
  - Voikko for Finnish language understanding
- MariaDB, MongoDB, (Apache Cassandra)
- LDAP based user management
  - OpenLDAP reference implementation
- SOSWE
  - Custom developed distributed micro-service workflow engine
  - Open source
  - Looking for and building micro-services
- Jasper Reports
- Piwik
- (however, need for some additional proprietary tools)

MAMK
University of Applied Sciences

Current status (and issues)

- Currently in Beta
- Implemeneting pilot tests with project partners
- Looking positive but …
- Fedora 3 issues
  - Performance and scalability (with batches and massive operations)
  - Complexity (configurations, content models)
  - Lack of transactions and multi-tenancy
  - Lack of knowledge (and docs, examples, up-to-date references)
- Middleware issues
  - Gsearch
  - Message queue persistence and keeping Solr in sync

MAMK
University of Applied Sciences



MAMK
University of Applied Sciences

Fedora 4

Key requirements

- Good design and simplicity (from developer point of view)
- REST API
- Performance upgrades
- Batch operations
- Transactions
- RDF and linked data support
- Powerful but simplified content modeling
- Multi-tenancy

What we can contribute

- Use cases and testing
- Java client development
- Promotion
- Project deliverables (once completed and decided licensing)

MAMK
University of Applied Sciences

Future development

Project scope

- Workflow engine and micro-services
- User experience upgrades
- F4 Java client
- API
- NoSQL storage for access
- Reporting and analytics

Future future development

- Personal archiving
- Productization and migration with commercial services
- Utilization with new industries

Open source tape management -- contact us

MAMK
University of Applied Sciences

Review

What we did in a nutshell

1. Design drivers and requirements identification
2. Data model design
3. Software review and analysis
4. Hand-crafted software to exploit Fedora and the best tools
5. Share and profit

MAMK
University of Applied Sciences

Links and deliverables

- Follow OSA - Open Source Archive blog and twitter
  - http://osarchive.wordpress.com/
  - @OSArchive

- OSA project final report in English will be available by end of 2014.
- Capture project summary is available in English.
  - Complete documentation in Finnish

- www.mamk.fi/osa (Finnish only)
- Ask anything: mikko.lampi@mamk.fi or @jotudin in Twitter