eHumanities - Data

**Project**
- Research data from the Digital Humanities

**Resources**
- Lexical information from the Alpine region (=pilot project), with approx 500,000 datasets
- Many relations between datasets (hasPart, isPartOf, hasPreviousVersion, hasNextVersion, ...)

**Metadata**
- rdUB = metadata schema which extends DataCite by new fields:
  - contentUrl
  - metadataUrl
  - checksum
  - ...

eHumanities - Infrastructure

Resources
• Fedora 5.1.1
• Apache Solr 8.1
• Project Blacklight 7.1

Tools and Integration
• Apache Camel 2.24.2
  • Components: XSLT, http4, Solr, CSV, XML, JSON-LD
eHumanities - Infrastructure
eHumanities - Infrastructure

**Features**

- Minted UUID and automated creation/administration of both container/binaries in Fedora
- Extensible: Ingest of new versions of VerbaAlpina binaries possible
- Transformation of files occurs before and after ingest. Heavy processing on Fedora and Camel
eHumanities - discussion

Challenges
• Scalability: limits of Fedora / Solr / Camel during ingest
• Concurrent operations in Fedora: what are the limits for concurrent GET / POST / PUT Requests / Event Message clogging?

Open questions
• Cluster vs Standalone installation (performance-wise)? (are Cluster installations of Fedora still available?) (Best tweaks for Fedora 5?)
• Impact of an eventual migration to Fedora 6 (API, data migration)?
eHumanities

Thank you!

fdm@ub.uni-muenchen.de
jaime.penagos@ub.uni-muenchen.de