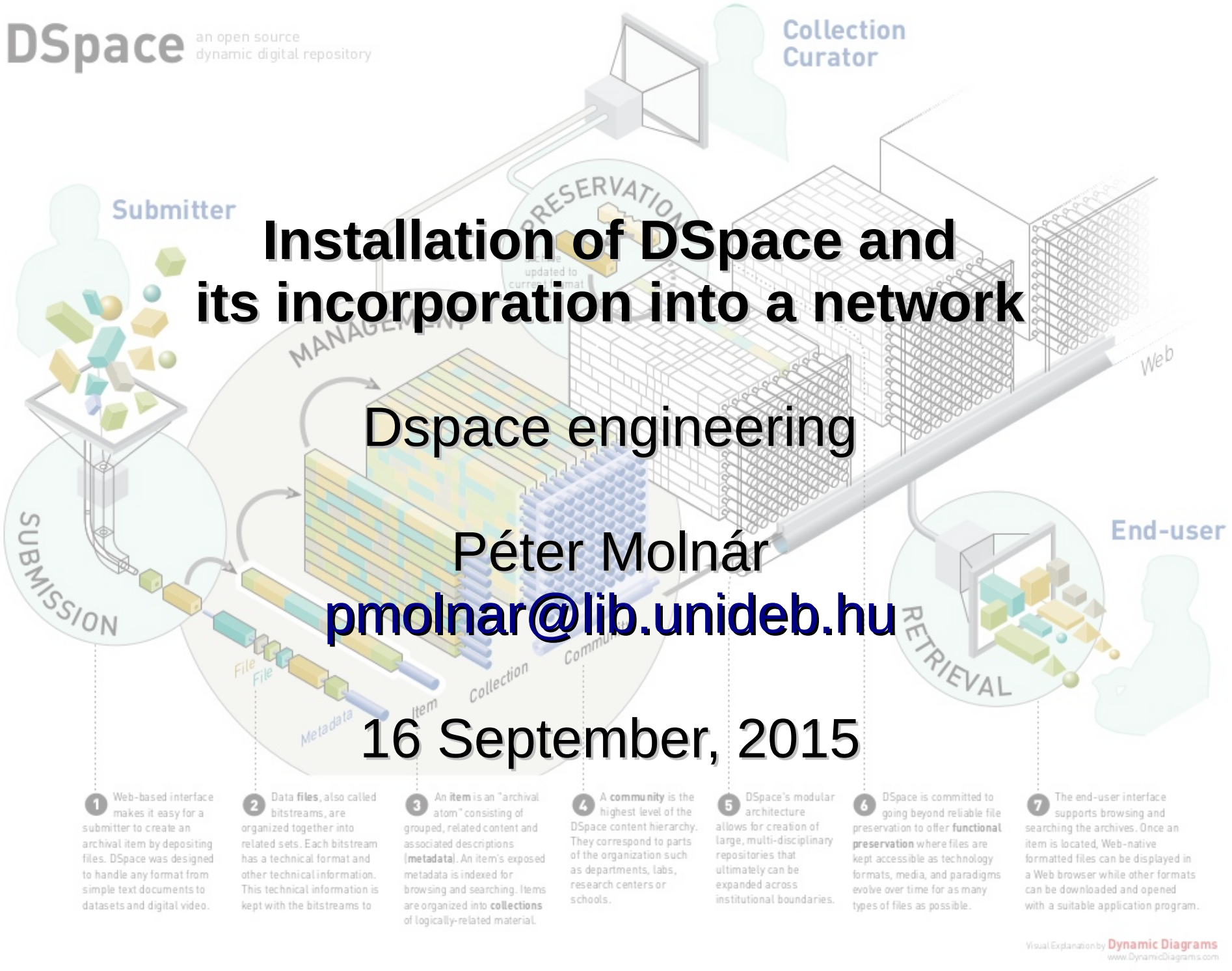


Installation of DSpace and its incorporation into a network

Dspace engineering

Péter Molnár
pmolnar@lib.unideb.hu

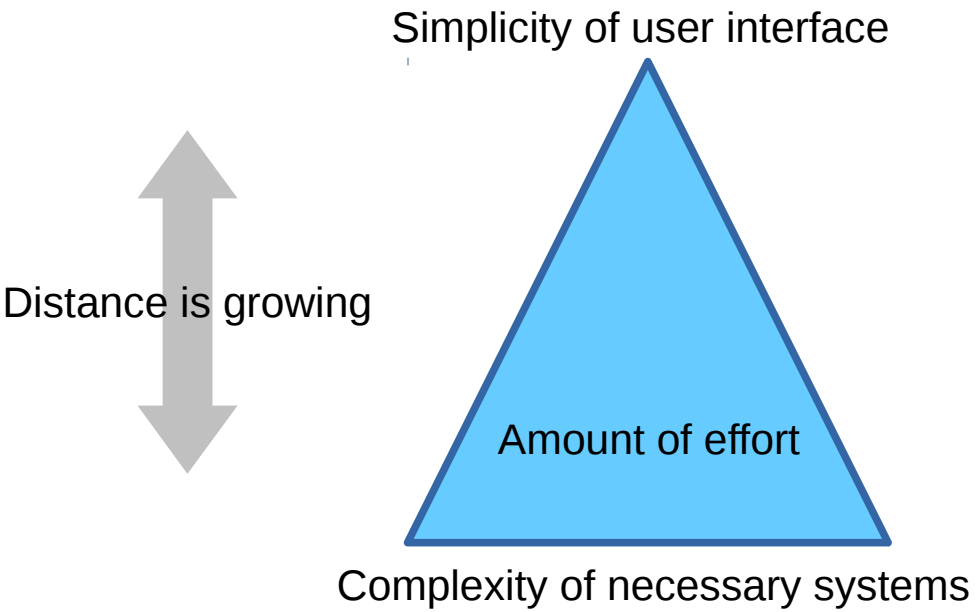
16 September, 2015

In this presentation

1. Computer system engineer's point of view
2. Conceptual overview of digital repositories
3. Long-term data storage problems
4. Reliable DSpace operating environment
5. Backup, security, disaster recovery
6. Integration to other systems
7. Quick installation guide

Users need

- User interface need to be simple
- Complexity of backend systems grows
- It needs more development work



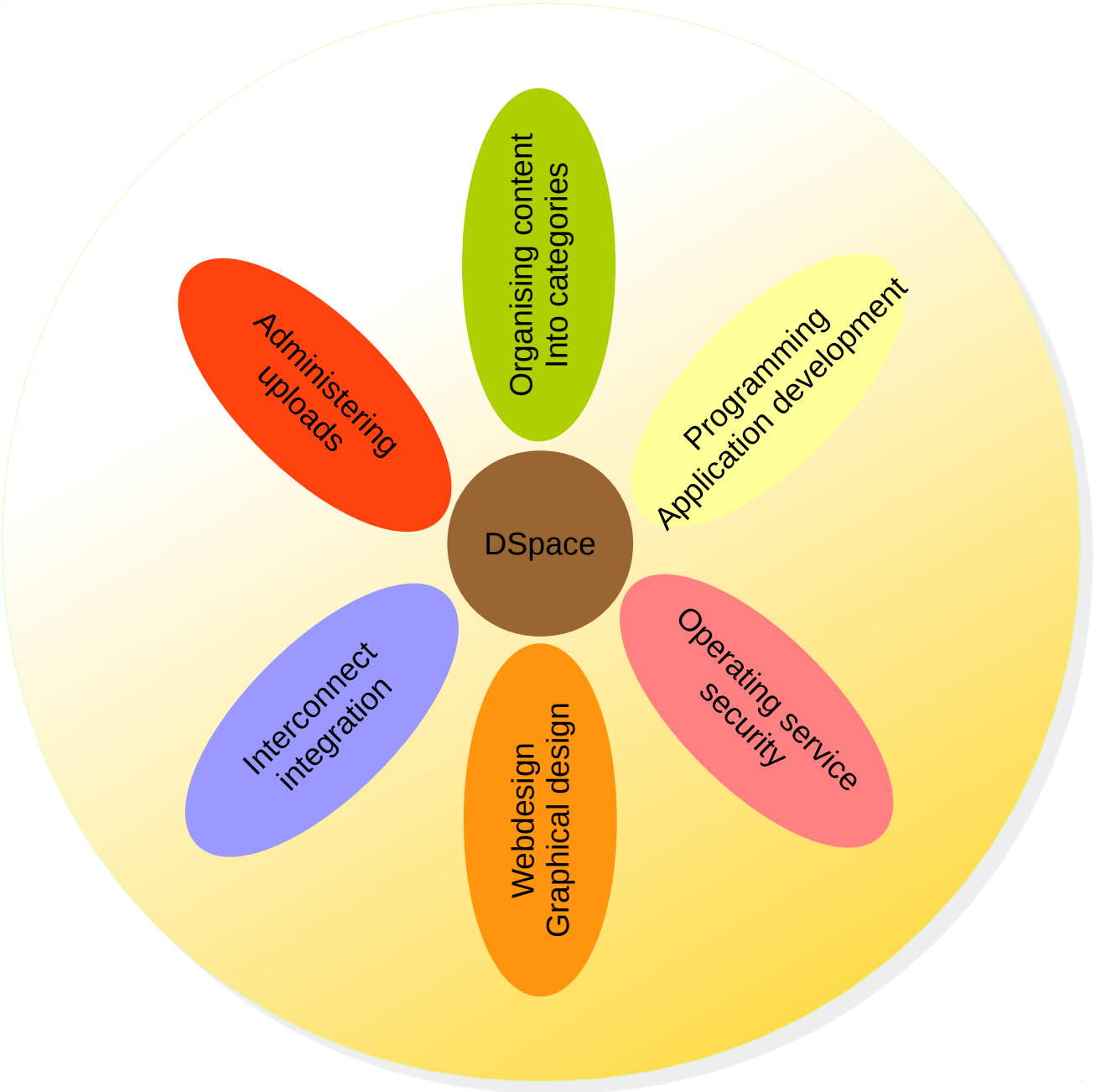
Digital archives

- Managing…
- … content (files)
- … metadata (database)
- … submit, upload (procedure)
- … access (procedure)
- … storage (maintain files on disks)
- … connection to other systems (integration)

What is DSpace?

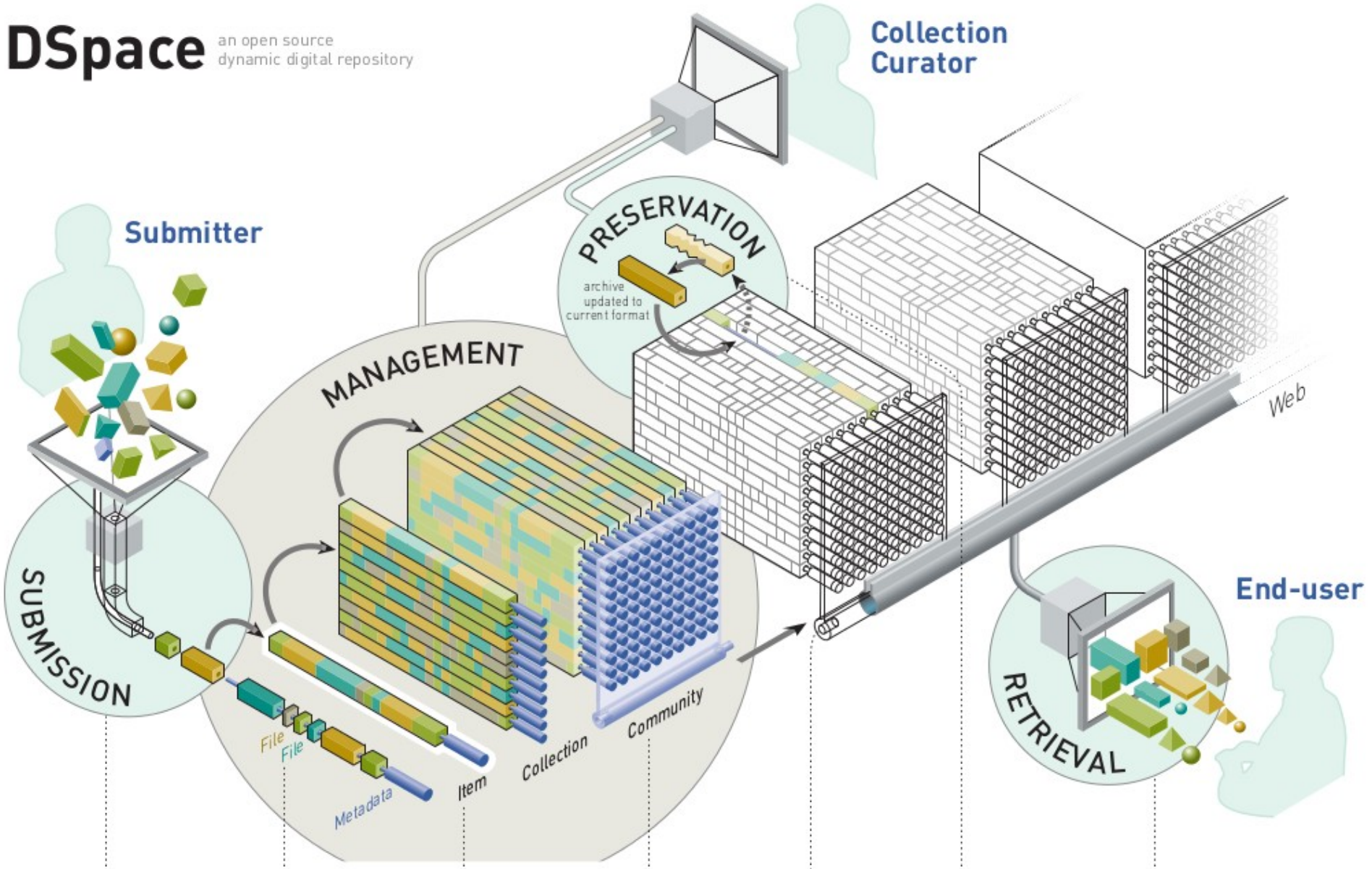
- Universal, digital object repository
- Search engine
- Access control manager
- Content upload workflow manager
- Open source software
- Tomcat based Java webapplication
- Database and filesystem based application
- The example in this presentation

Tasks with a digital repository



DSpace

an open source
dynamic digital repository



1. Web-based interface makes it easy for a submitter to create an archival item by depositing files. DSpace was designed to handle any format from simple text documents to datasets and digital video.

2. Data **files**, also called bitstreams, are organized together into related sets. Each bitstream has a technical format and other technical information. This technical information is kept with the bitstreams to

3. An **item** is an "archival atom" consisting of grouped, related content and associated descriptions (**metadata**). An item's exposed metadata is indexed for browsing and searching. Items are organized into **collections** of logically-related material.

4. A **community** is the highest level of the DSpace content hierarchy. They correspond to parts of the organization such as departments, labs, research centers or schools.

5. DSpace's modular architecture allows for creation of large, multi-disciplinary repositories that ultimately can be expanded across institutional boundaries.

6. DSpace is committed to going beyond reliable file preservation to offer **functional preservation** where files are kept accessible as technology formats, media, and paradigms evolve over time for as many types of files as possible.

7. The end-user interface supports browsing and searching the archives. Once an item is located, Web-native formatted files can be displayed in a Web browser while other formats can be downloaded and opened with a suitable application program.

Visual Explanation by **Dynamic Diagrams**
www.DynamicDiagrams.com

Long-term storage

Questions about long-term storage

- File formats
- Data conversion
- Persistent access methods
- http://spectrum.ieee.org/computing/hardware/external-bits

Long-term storage

- File formats
  - File formats change (MS Word 1.1a – MS Word 2016)
  - Persistence
- Software that can read files
- Operating system for this software
- Hardware for this operating system

Long-term storage

- Data conversion
  - Old file formats need to be converted
  - Conversion of proprietary formats is difficult
  - A lot of manual work
  - Clear text and binary formats
- Data storage mediums
  - Tape, floppy, hard disk, optical disc, cloud

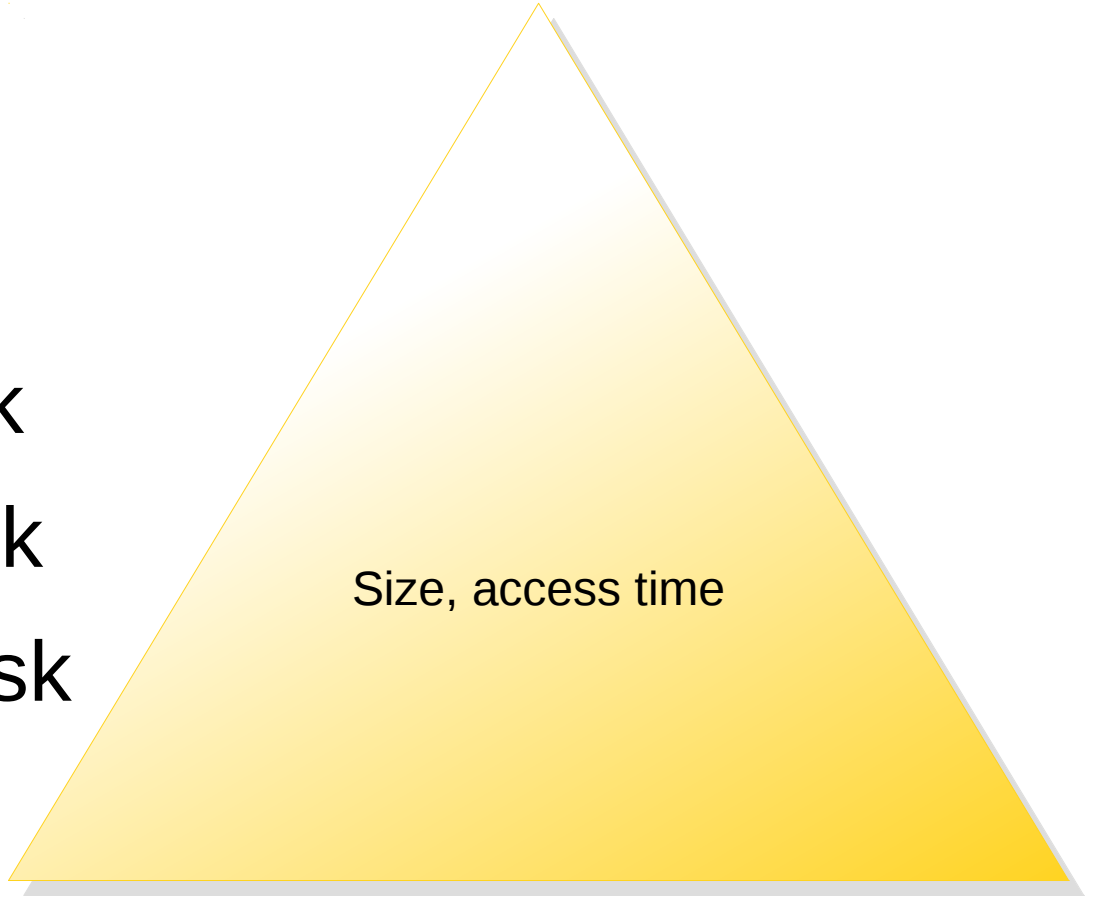
Long-term storage

Persistent access methods

- An object URL on the web must be constant
  - https://repo.library.edu/object/xxxx
  - ~~https://dspace.library.edu/object/xxxx~~
- And redirects for many years
  - https://repo.new-name-of-the-library.edu/object/xxxx
- Design for future
-

Storage tiering

- Memory
- SSD
- Fast hard disk
- Slow hard disk
- Half-online disk
- Tape
- ~~CD, DVD, Blu-ray~~ not for digital repository



Cloud and digital repositories

Responsibility

- Data assets are expensive
- Librarians are responsible for operating a library
- Public cloud providers are not librarians …
- … but they offer good services.
- Can we move frontend or storage to the cloud?

Hardware recommendation

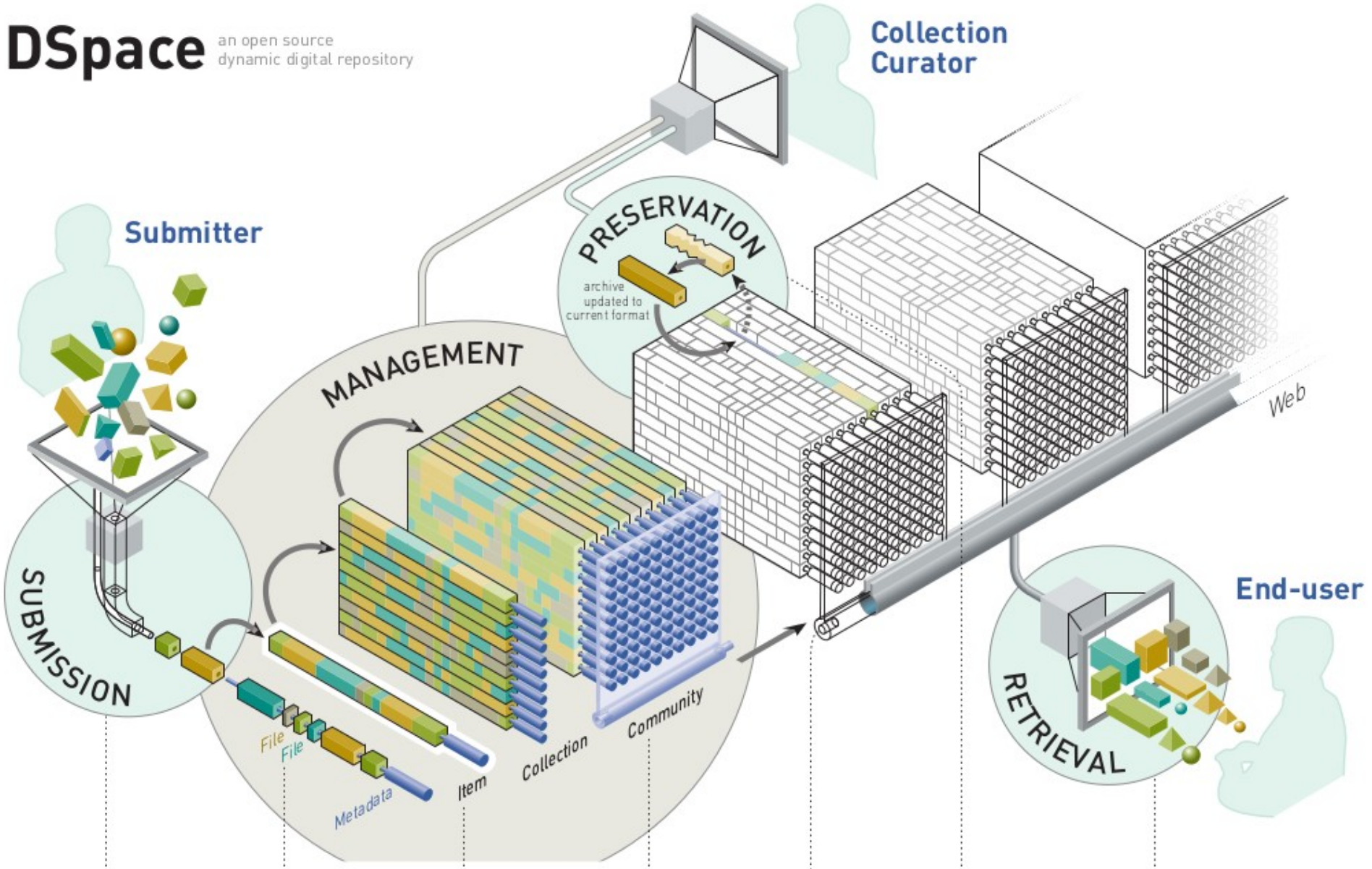
- Real server hardware
  - Multiple, redundant disc (RAID)
  - Error checking system memory (ECC)
- Not serverized desktop PC
  - Possibly power problems
  - Possibly data consistency problems
- Database and assetstore: sensitive information
  - Enterprise grade disk or SSD

Hardware environment recommendation

- Server room
- Air conditioning
- Fire-extinguisher
- No wind
- No water
- Gb speed network link

DSpace

an open source
dynamic digital repository



1. Web-based interface makes it easy for a submitter to create an archival item by depositing files. DSpace was designed to handle any format from simple text documents to datasets and digital video.

2. Data **files**, also called bitstreams, are organized together into related sets. Each bitstream has a technical format and other technical information. This technical information is kept with the bitstreams to

3. An **item** is an "archival atom" consisting of grouped, related content and associated descriptions (**metadata**). An item's exposed metadata is indexed for browsing and searching. Items are organized into **collections** of logically-related material.

4. A **community** is the highest level of the DSpace content hierarchy. They correspond to parts of the organization such as departments, labs, research centers or schools.

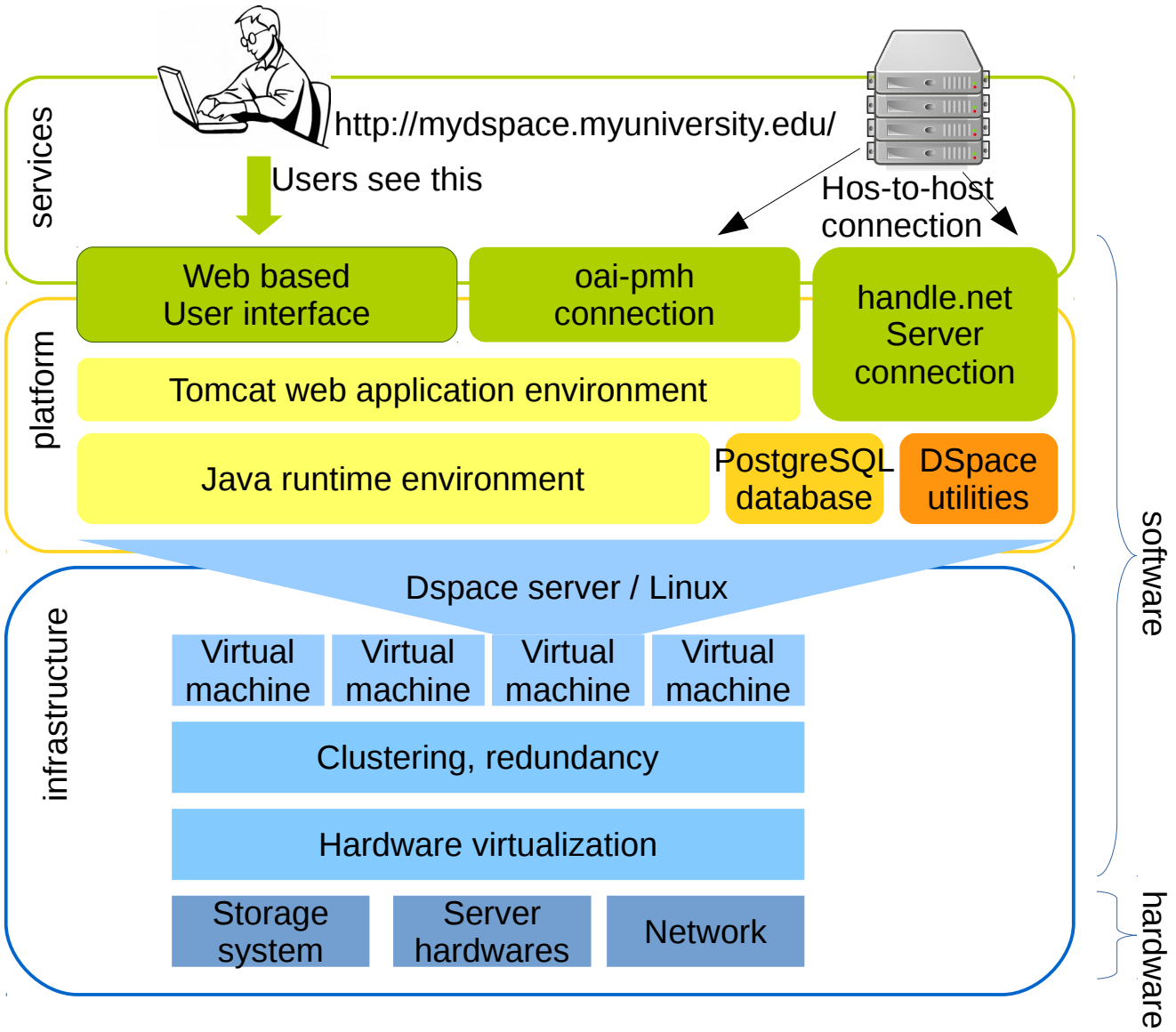
5. DSpace's modular architecture allows for creation of large, multi-disciplinary repositories that ultimately can be expanded across institutional boundaries.

6. DSpace is committed to going beyond reliable file preservation to offer **functional preservation** where files are kept accessible as technology formats, media, and paradigms evolve over time for as many types of files as possible.

7. The end-user interface supports browsing and searching the archives. Once an item is located, Web-native formatted files can be displayed in a Web browser while other formats can be downloaded and opened with a suitable application program.

Visual Explanation by **Dynamic Diagrams**
www.DynamicDiagrams.com

System architecture



Backup

- Metada, database relatively small
- Assetstore is large (even bigger with videos)
- Regular (daily) differential rsync
- Preservation of full backup monthly, yearly
- Encrypted backup in a public cloud
- Disk space requirements: min. assetstore size (+ online service)

Security

- Web access over HTTPS (force it)
- Buy a commercial certificate authority signed certificate
- Avoid unencrypted trafic (username, password)
- Frequent updating (all components: DSpace, Java, Tomcat, OS, utils)
- Virus check of administrators PCs also
- Avoid full-control admin accounts

Security

- Auto attacks
- Overload attacks
- Plagiarism: illegal copy – security
- Jump to university study information system
- Data asset is expensive
- Encryption in cloud backup

Disaster recovery

- Cause of disaster:
  - Human,
  - Technical,
  - Vis-major
- Several backup copies, geo-distribution
- Disaster recovery training: everything works?

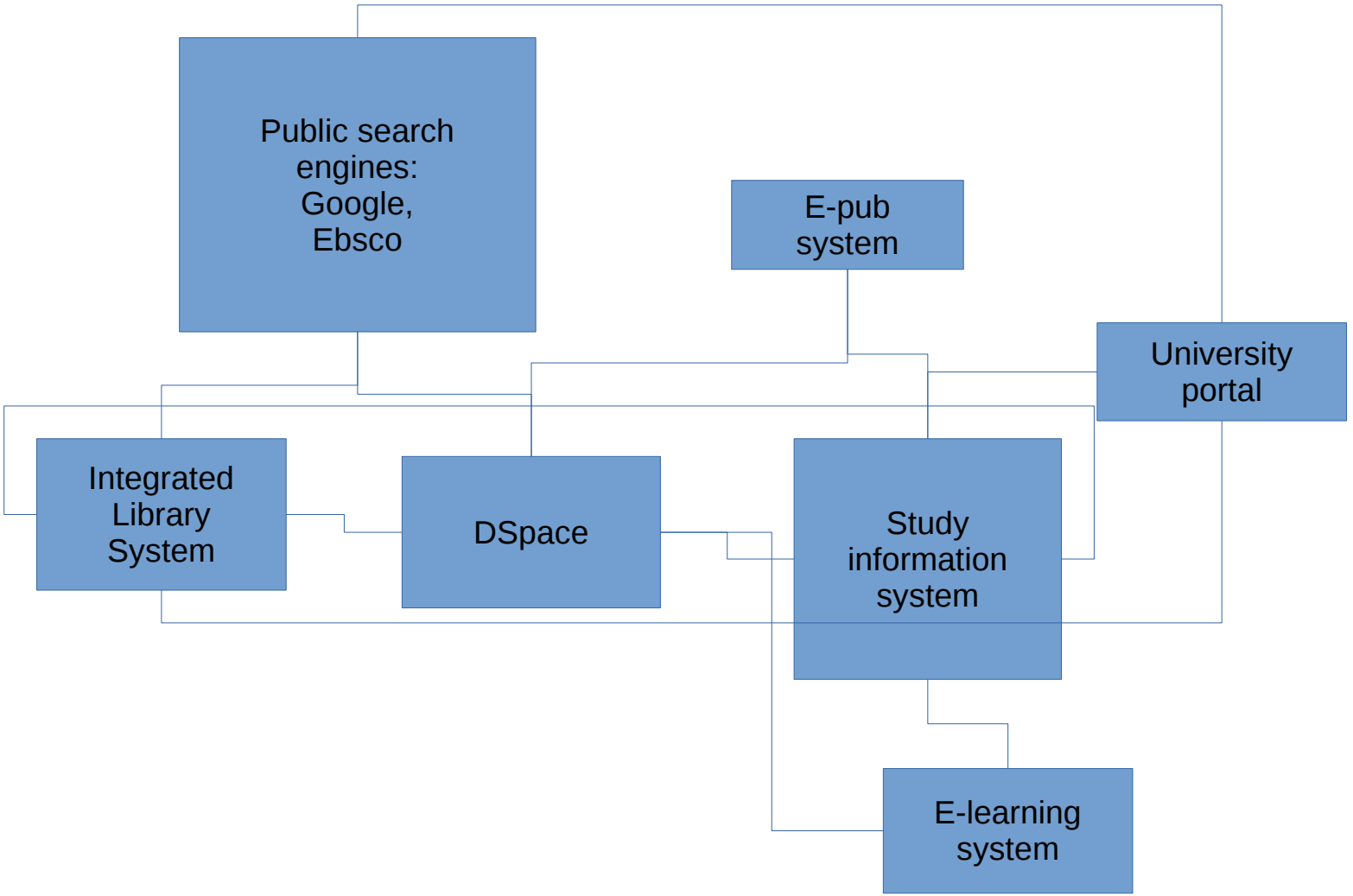
Integration

- Simple integration possibilities
  - Using central (LDAP, ActiveDirectory) login server
  - OAI data harvesting: search servers can harvest (download) metadatas of the repository
- Advanced integration possibilities
  - Institutional Single Sign-On (SSO)
  - E-journal softwares (SWORD)
  - Plagiarism check

Customization

- Most advanced step
- Graphical design
- Behavioural changes
  - Submit form and process
  - DeenkStatistics
- Viewers and players
- Bugfixes needs developers
- Source code copyrights

Integration



Software recommendation

- Linux
- PostgreSQL
- Java 7
- Tomcat 7
- For building and installation
  - JDK
  - Ant
  - Maven
-

Installing, building

Download Tomcat and DSpace

- http://tomcat.apache.org/download-70.cgi
- http://www.dspace.org/latest-release

Prepare on Debian as root

```
apt-get install ant ffmpeg imagemagick maven openjdk-7-jdk postgresql

useradd -d /var/opt/dspace -m -r dspace

mkdir -p /opt/dspace

chown -R dspace:dspace /opt/dspace

su - postgres -c "createuser -D -R -P -S dspace"

su - postgres -c "createdb -E UNICODE -O dspace dspace"
```

Installing, building

Setting up Tomcat

cd /opt

tar -xf apache-tomcat-7.0.XX.tar.gz

chown -R dspace:dspace /opt/apache-tomcat-7.0.XX

Edit: *<u>/opt/apache-tomcat-7.0.XX/conf/server.xml</u>*

```
<Connector port="8080
        maxThreads="150"
        minSpareThreads="25"
        maxSpareThreads="75"
        enableLookups="false"
        redirectPort="8443"
        acceptCount="100"
        connectionTimeout="20000"
        disableUploadTimeout="true"
        URIEncoding="UTF-8"/>
```

Installing, building

Prepare DSpace webapps config in Tomcat

```
mkdir -p /opt/apache-tomcat-7.0.XX/conf/Catalina/localhost
cd /opt/apache-tomcat-7.0.XX/conf/Catalina/localhost
```

Create *xmlui.xml*

```
<?xml version='1.0'?>
<Context
   docBase="/opt/dspace/webapps/xmlui"
   debug="0"
   reloadable="true"
   cachingAllowed="false"
   allowLinking="true"/>
```

Create *solr.xml*

```
<?xml version='1.0'?>
<Context
   docBase="/opt/dspace/webapps/solr"
   debug="0"
   reloadable="true"
   cachingAllowed="false"
   allowLinking="true"/>
```

Installing, building

Installing DSpace

```
tar -xf dspace-X.X-src-release.tar.gz

cd dspace-X.X-src-release

Edit build.properties: dspace.install.dir, dspace.hostname,
dspace.baseUrl, dspace.name, default.language, db.password

mvn package

ant fresh_install

bin/dspace create-administrator
```

Start Tomcat

```
su – dspace

/opt/apache-tomcat-7.0.XX/bin/startup.sh
```

Installing, building

Open DSpace in a browser and login as administrator

http://<dspace.baseUrl>/xmlui