Implementation of the information system BExIS 2 at the UFZ: Quality control, retrieval and sharing of [biodiversity] data

Dr. Mark Frenzel

Helmholtz Centre for Environmental Research - UFZ

Best practice elements and goals in **data management**

AIM: Reuse of data!

- Matter of **attitude** of people
  - Recognition of importance of data management
    - Top down and Bottom up!
  - Availability and **Open access** to data
- Defined **workflows** (acquisition-quality control-storage-publication)
- **Documentation** and selection of relevant data sets
  - Meta data standard for data description
- **IT issues**
  - Thesauri (controlled vocabulary)
  - Persistent storage / hardware
  - Magic tools: software solutions

Best practice elements and goals in **data management**

- Final step: DOI Data **publication** of **relevant** data sets

Linking data to publications and people

↺ Feed back on willingness of data providers

Smiling data creators

Smiling users

The issue of **biodiversity data**

Mostly person-generated data!

- Heterogeneity of data
- Logic of ecologists related to data (different from IT people)
  - Ecology: data based on spreadsheets ⇨ Data base?!

Quality control

- Plausibility tests
  - Expert knowledge
  - Software (e.g. occurrence of species A at location B plausible?)
- Technical consistency
  - Correct data types
  - Correct cell entries

BExIS - a generic data management **tool for biodiversity data**

Biodiversity **Ex**ploratories **I**nformation **S**ystem

- **BExIS 1**: Development started with DFG project Biodiversity Exploratories (2006) ⇨ information management system, project data base
  - **Instances**: DFG Biodiversity Exploratories, DFG Jena Experiment, DFG Research Group: Kilimanjaro, DFG Collaborative Research Centre 990: Ecological and Socioeconomic Functions of Tropical Lowland Rainforest Transformation Systems (Sumatra, Indonesia)
- **BExIS 2** (DFG-Project): generic open source information system for biodiversity data (funded until 2017; http://bexis2.uni-jena.de; live demo; download BExIS)
  - **Instances**: iDiv (DFG), AquaDiva (DFG), UFZ

BExIS: basic features

- **Features**
  - **Access**: free, as generic tool not restricted to biodiversity data!
  - **Import** of structured (spreadsheet-based) and unstructured data (e.g. images)
  - Internal **table-to-database** conversion
  - Data type **consistency check**
  - **Metadata** (import structures as xsd = xml schema definition)
  - **Export** (csv, xlsx)
  - Administration of **admission rights**
  - **Modular architecture** (data planning, data collection, data dissemination, data discovery, system administration)

BExIS: basic advantages

- Ideal for (large) projects and groups
  - all data including metadata are at **one place** (data base mangement in background)
  - **Web** interface
  - Individual data **access** management
  - Data base: even **search** within primary data
  - Ingests **all kind of data**
  - Dataset **versioning**

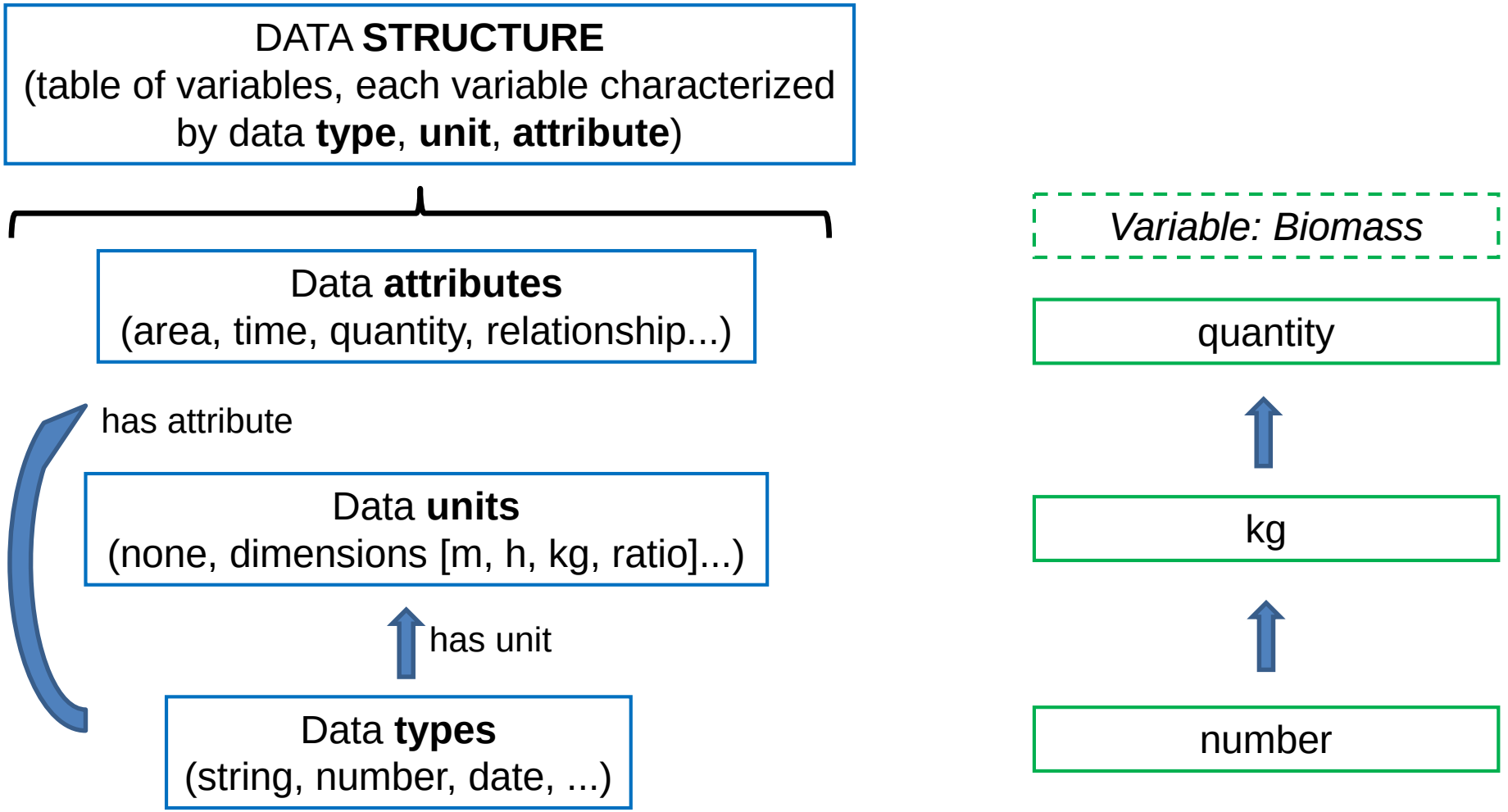
- Close interaction users ⇔ developers in project runtime
  - User and developer **conference June 9-10, 2016 in Jena** (Germany)

For IT administrators: Running BExIS

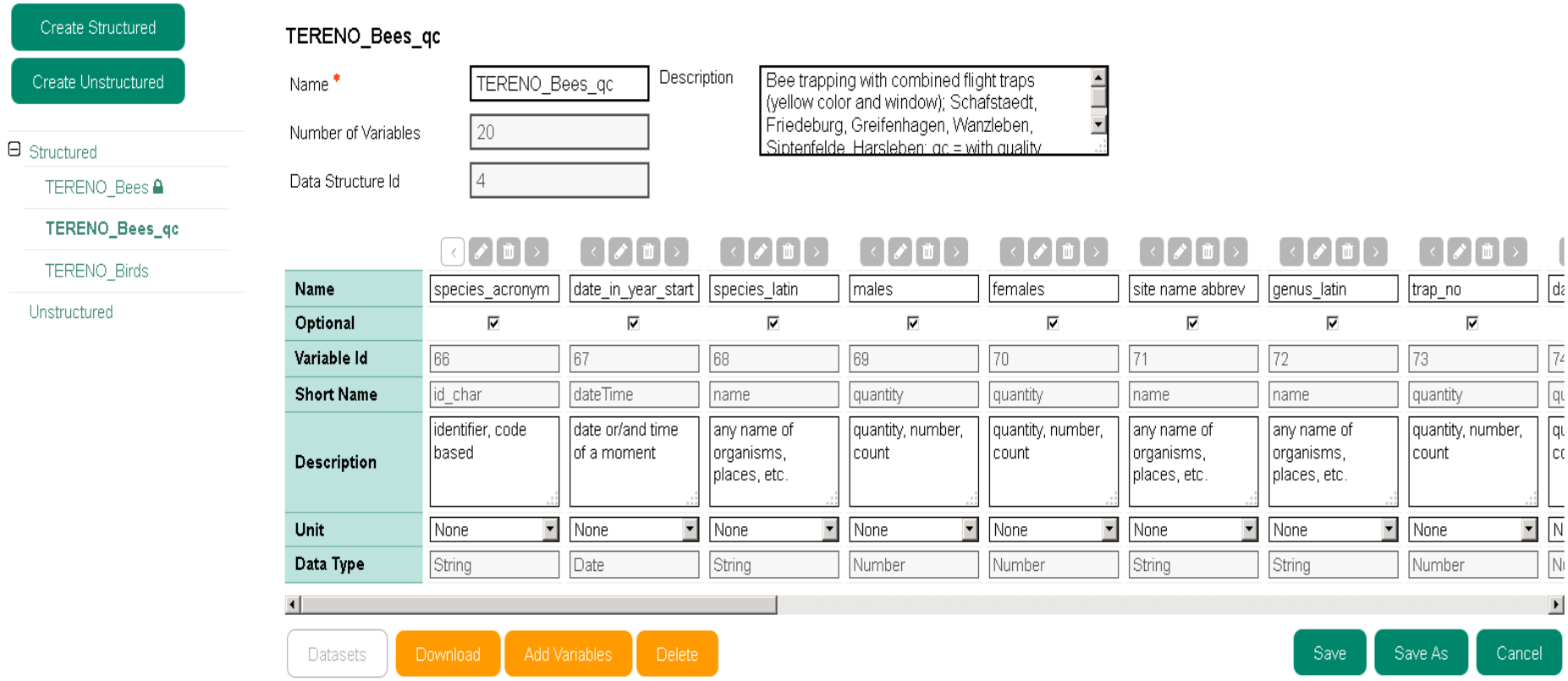
- Installation requirements
  - PostgreSQL or IBM DB2 Express-C
  - .NET Framework 4.5.2
  - Internet Information Service (IIS; Microsoft web server)

- UFZ instance
  - Virtual machine in DMZ – DeMilitarized Zone; outside firewall
  - Connected to LDAP (Lightweight Directory Access Protocol) ⇨easy login for UFZ users
  - accessible as web application within intranet UFZ (bexis.ufz.de)
  - https access for outside world possible

Getting organized by software

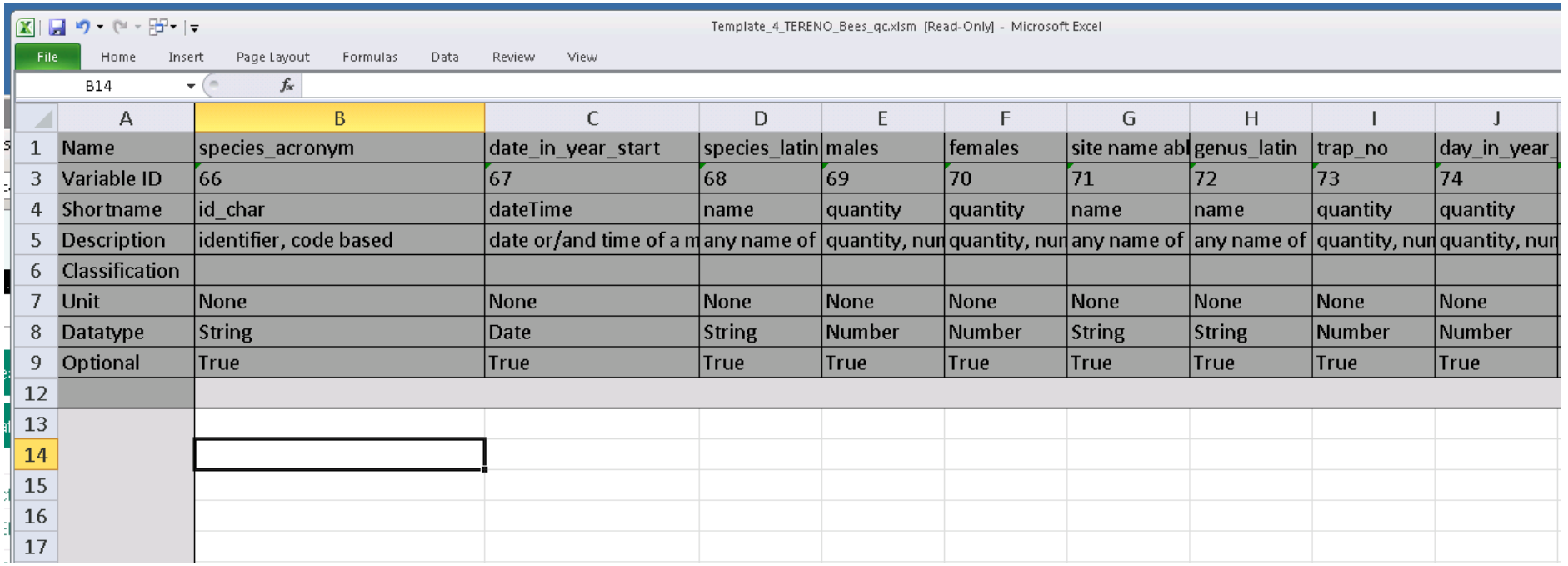


Data structure ⇨ download Excel template



Excel template (xlsm)

Template with complete data structure entries and **makro** running in the background



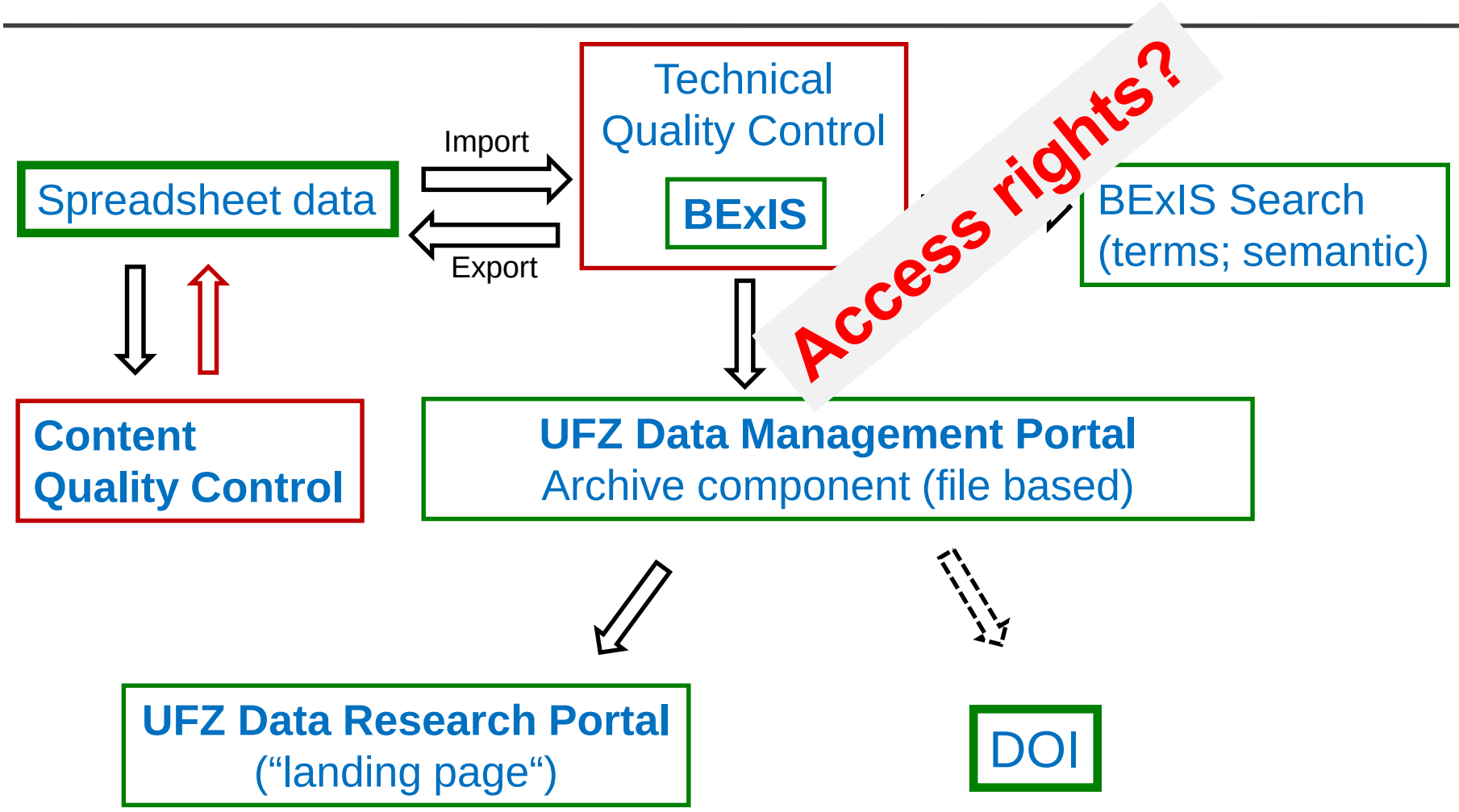
	A	B	C	D	E	F	G	H	I	J
1	Name	species_acronym	date_in_year_start	species_latin	males	females	site name abl	genus_latin	trap_no	day_in_year_
3	Variable ID	66	67	68	69	70	71	72	73	74
4	Shortname	id_char	dateTime	name	quantity	quantity	name	name	quantity	quantity
5	Description	identifier, code based	date or/and time of a m	any name of	quantity, nun	quantity, nun	any name of	any name of	quantity, nun	quantity, nun
6	Classification									
7	Unit	None	None	None	None	None	None	None	None	None
8	Datatype	String	Date	String	Number	Number	String	String	Number	Number
9	Optional	True	True	True	True	True	True	True	True	True
12										
13										
14										
15										
16										
17										

Excel template (xlsm)

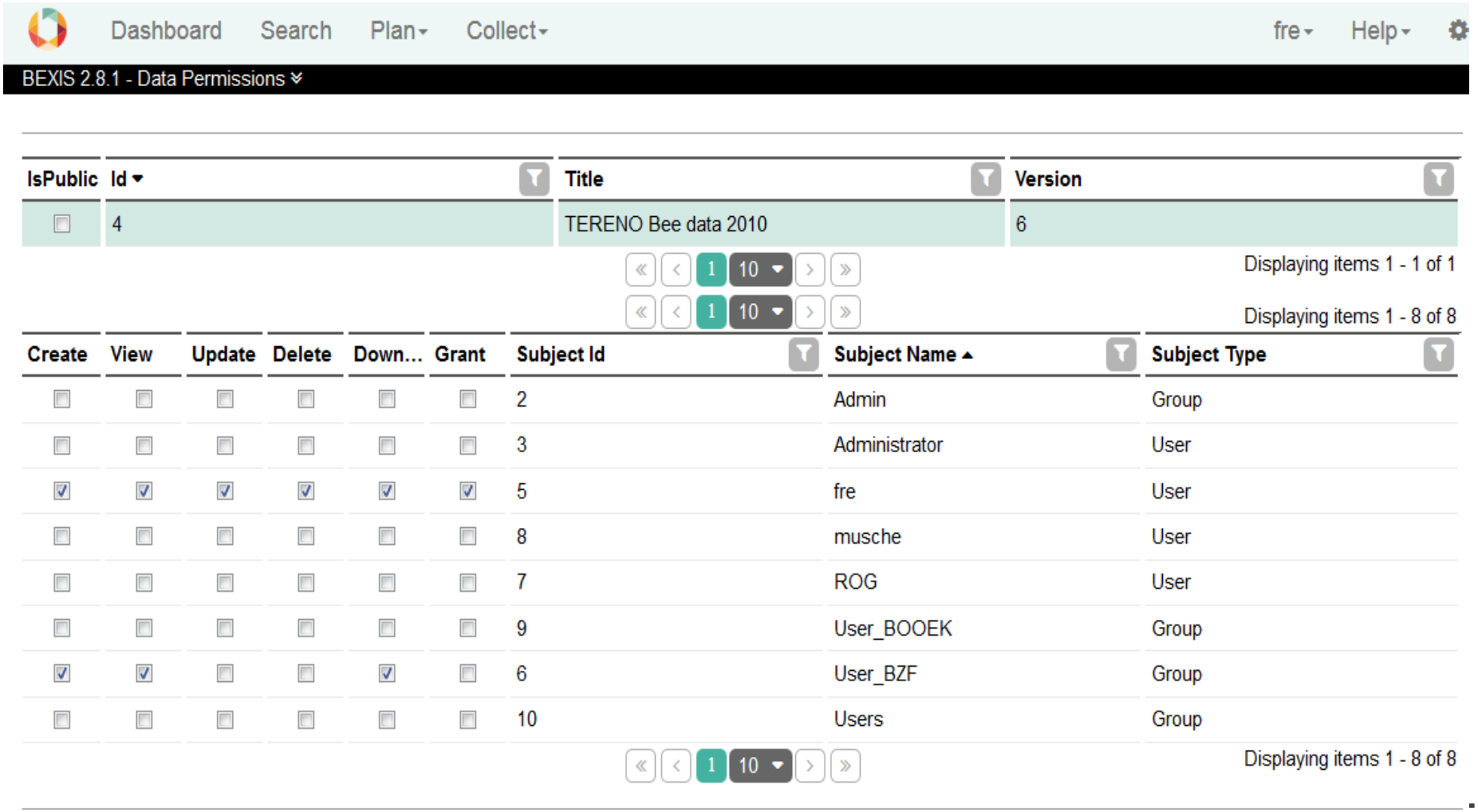
- Copy & paste your data in the template
- Data type consistency check ⇨ example: “test“ is no number and thus indicated by the red cell

F	G	H	I	J	K	L
gen_spec_latin	males	females	site name ab	trap_no	week_start	date_in_yea
78	87	88	71	73	82	67
name	quantity	quantity	name	quantity	quantity	dateTime
latin species name	quantity, nur	quantity, nur	observation	number of th	number of w	date when a
None	None	None	None	None	None	None
String	Number	Number	String	Number	Number	Date
True	True	True	True	True	True	True
Andrena flavipes	22	2	FBG	test	21	02.05.2010
Andrena haemorrhc	0	1	FBG	1	21	02.05.2010
Andrena helvola	0	1	FBG	1	21	02.05.2010
Andrena minutula	0	1	FBG	1	21	02.05.2010
Andrena nigroaenea	4	5	FBG	1	21	02.05.2010
Andrena propinqua	0	1	FBG	1	21	02.05.2010
Andrena proxima	1	0	FBG	1	21	02.05.2010
Andrena scotica	0	1	FBG	1	21	02.05.2010
Andrena strohmella	0	2	FBG	1	21	02.05.2010
Andrena synadelph	2	0	FBG	1	21	02.05.2010

Workflow for biodiversity data at UFZ



Manage users | groups | features | **data permissions**



IsPublic	Id	Title	Version
☐	4	TERENO Bee data 2010	6

« ‹ 1 10 › » Displaying items 1 - 1 of 1

« ‹ 1 10 › » Displaying items 1 - 8 of 8

Create	View	Update	Delete	Down...	Grant	Subject Id	Subject Name	Subject Type
☐	☐	☐	☐	☐	☐	2	Admin	Group
☐	☐	☐	☐	☐	☐	3	Administrator	User
☑	☑	☑	☑	☑	☑	5	fre	User
☐	☐	☐	☐	☐	☐	8	musche	User
☐	☐	☐	☐	☐	☐	7	ROG	User
☐	☐	☐	☐	☐	☐	9	User_BOOEK	Group
☑	☑	☐	☐	☑	☐	6	User_BZF	Group
☐	☐	☐	☐	☐	☐	10	Users	Group

« ‹ 1 10 › » Displaying items 1 - 8 of 8

Larger context: www.**gfbio**.org

German **F**ederation for **Bio**logical Data (GFBio; DFG project; BExIS is a component)

"sustainable, service oriented, *national data infrastructure* facilitating data *sharing* and stimulating data intensive science in the fields of *biological and environmental* research"

- **Data focus**: genome data, ecological and environmental data, collection related data
- **Coverage**: full life cycle of research data ⇨ field or real time data acquisition ⇨ long term archiving ⇨ **publication** ⇨ re-analysis and re-use

From data management to **DOI for data sets**

DOI = Digital Object Identifier

BExIS ⇨ important step towards DOI quality of data sets

Why DOI for data sets?

- **Credits** to data producers / owners
- **Persistent** identifiers, persistant storage
- Standardised **metadata**
- Increasing requirement from **publishers**
- Easy access *via* individual **landing page** (url) for each data set

From data management to **DOI for data sets**

One option ⇨ PANGAEA (www.pangaea.de; publication agent for dataset DOI)

Features of PANGAEA

- Jira **ticket system** for data submission and documentation
- **Editorial system** (4D client)
- Structured data splitted to **database**
- **Ontologies** behind
- **Database + Ontology = Data warehouse** ⇨ essential for **reuse** and new combination of related datasets!

Link to exemplary landing page in PANGAEA