Documentation for the Individual Project for ISY Course

"Version control of institutional data with MongoDB"

Project By: Vineet Sharma

Supervised by: Cord Wiljes

1. Introduction to the Project

In the fourth Semester of the study of "Informatik – Intelligent Systems" the project Data Mining on Institutional Data was initiated by Cord Wiljes. The project was developed by Vineet Vikram Sharma.

There was an official description for the project at the beginning and according to this theoretical description different background work regarding the choice of technology and extraction and storage of data was carried out. Python and MongoDB was chosen to extract information from the available data.

The institutional data we used contains details of list of persons, data about individual persons and publication data. The total no of xml files were roughly around 10,000.

2. Technology Used

I decided to go with Python and MongoDB because I was keen to continue working on the same dataset and progress to find new results/findings from the data. So, MongoDB as it was a NOSQL database and was a good choice as the data was not completely structured. Being able to handle xml and json formats natively in a range of NoSQL databases lessened the amount of code we have to convert from the source data format to the format that needs storing.

Python gave the advantage to quickly integrate with MongoDB and parse the data in xml. For personal data, we parsed through each node but for publication data, we took data which were relevant and related to University knowledge base. Xml dom minidom and xpath were used for parsing.

3. Overview over the System

The data were extracted from the xml files from http://ekvv.uni-bielefeld.de/pers_publ/ server and stored in the local database according to the different tags in the response xml file along with the timestamp when the request is sent to the server.

The following methods of backup were used in the dataset:

Full Backup: contains the whole dataset present in the current copy.

Incremental Backup: contains only the data subset that has changed since the preceding backup copy.

Differential Backup: saves the difference in the data since the last full backup.

The data were stored were archived and stored into following four tables:

Personal First Copy (personal_first_copy): It is the set of data when it was first extracted.

Personal (personal): It is a set of data which are created when a request is send to scan through the xml files from the server. It is the mirror of the datasets present in the latest version of the server.

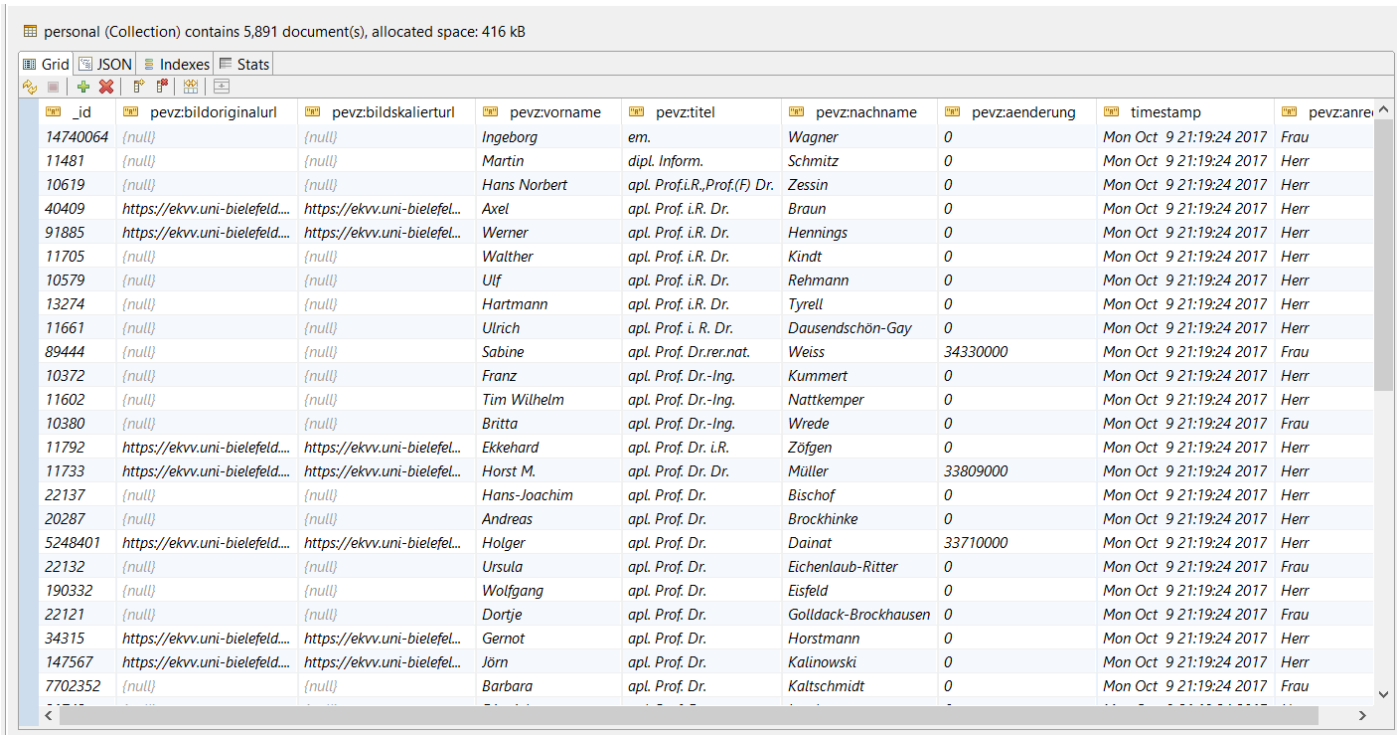
Personal Archive (personal_arcvive): It is the set of data which incremental backup of all the dataset scrapped during the server requests. It contains all the dataset that appeared previously and currently in the server along with the timestamp.

Personal Archive Differential (personal_arcvive_differential): It contains all the fields which are changed in the current copy in reference to the first copy along with the timestamp when the change was discovered.

The format of the xml files used. (PersonDetailxx.xml)

```
<?xml version="1.0" encoding="UTF-8"?>
<pevz:kontakte personid="42996633" xmlns:pevz="http://ekvv.uni-bielefeld.de/pers_publ/">
    <pevz:kontakt id="42996634">
        <pevz:einrichtung id="2685608">
            <pevz:name>Fakultät für Biologie / Biologiedidaktik (Botanik und Zellbiologie)</pevz:name>
            <pevz:name_en>Faculty of Biology / Biologiedidaktik (Botanik und Zellbiologie)</pevz:name_en>
        </pevz:einrichtung>
        <pevz:funktion></pevz:funktion>
        <pevz:email>t.becker@uni-bielefeld.de</pevz:email>
        <pevz:email_verschleiert>&amp;#116;.be&amp;#99;ker&amp;#64;uni&amp;#45;bie&amp;#108;efe&amp;#108;d
        <pevz:url></pevz:url>
        <pevz:telefon>5683</pevz:telefon>
        <pevz:fax></pevz:fax>
        <pevz:telefon_sekretariat></pevz:telefon_sekretariat>
        <pevz:raum>UHG W2-226</pevz:raum>
        <pevz:sprechzeiten></pevz:sprechzeiten>
        <pevz:profs>nein</pevz:profs>
    </pevz:kontakt>
</pevz:kontakte>
```

Structure of the personal table in MongoDB

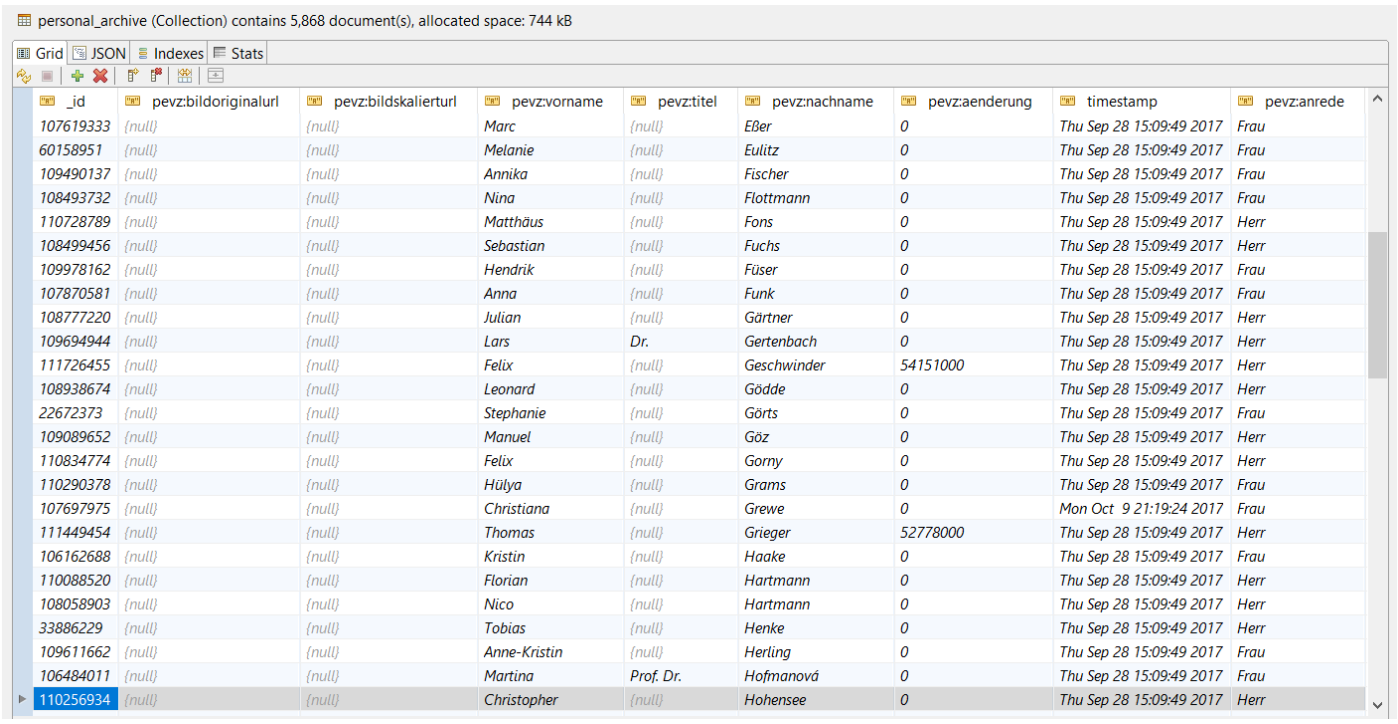


personal (Collection) contains 5,891 document(s), allocated space: 416 kB

Grid | JSON | Indexes | Stats

_id	pevz:bildoriginalurl	pevz:bildskalierturl	pevz:vorname	pevz:titel	pevz:nachname	pevz:aenderung	timestamp	pevz:anre
14740064	{null}	{null}	Ingeborg	em.	Wagner	0	Mon Oct 9 21:19:24 2017	Frau
11481	{null}	{null}	Martin	dipl. Inform.	Schmitz	0	Mon Oct 9 21:19:24 2017	Herr
10619	{null}	{null}	Hans Norbert	apl. Prof.i.R.,Prof.(F) Dr.	Zessin	0	Mon Oct 9 21:19:24 2017	Herr
40409	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Axel	apl. Prof. i.R. Dr.	Braun	0	Mon Oct 9 21:19:24 2017	Herr
91885	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Werner	apl. Prof. i.R. Dr.	Hennings	0	Mon Oct 9 21:19:24 2017	Herr
11705	{null}	{null}	Walther	apl. Prof. i.R. Dr.	Kindt	0	Mon Oct 9 21:19:24 2017	Herr
10579	{null}	{null}	Ulf	apl. Prof. i.R. Dr.	Rehmann	0	Mon Oct 9 21:19:24 2017	Herr
13274	{null}	{null}	Hartmann	apl. Prof. i.R. Dr.	Tyrell	0	Mon Oct 9 21:19:24 2017	Herr
11661	{null}	{null}	Ulrich	apl. Prof. i. R. Dr.	Dausendschön-Gay	0	Mon Oct 9 21:19:24 2017	Herr
89444	{null}	{null}	Sabine	apl. Prof. Dr.rer.nat.	Weiss	34330000	Mon Oct 9 21:19:24 2017	Frau
10372	{null}	{null}	Franz	apl. Prof. Dr.-Ing.	Kummert	0	Mon Oct 9 21:19:24 2017	Herr
11602	{null}	{null}	Tim Wilhelm	apl. Prof. Dr.-Ing.	Nattkemper	0	Mon Oct 9 21:19:24 2017	Herr
10380	{null}	{null}	Britta	apl. Prof. Dr.-Ing.	Wrede	0	Mon Oct 9 21:19:24 2017	Frau
11792	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Ekkehard	apl. Prof. Dr. i.R.	Zöfgen	0	Mon Oct 9 21:19:24 2017	Herr
11733	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Horst M.	apl. Prof. Dr. Dr.	Müller	33809000	Mon Oct 9 21:19:24 2017	Herr
22137	{null}	{null}	Hans-Joachim	apl. Prof. Dr.	Bischof	0	Mon Oct 9 21:19:24 2017	Herr
20287	{null}	{null}	Andreas	apl. Prof. Dr.	Brockhinke	0	Mon Oct 9 21:19:24 2017	Herr
5248401	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Holger	apl. Prof. Dr.	Dainat	33710000	Mon Oct 9 21:19:24 2017	Herr
22132	{null}	{null}	Ursula	apl. Prof. Dr.	Eichenlaub-Ritter	0	Mon Oct 9 21:19:24 2017	Frau
190332	{null}	{null}	Wolfgang	apl. Prof. Dr.	Eisfeld	0	Mon Oct 9 21:19:24 2017	Herr
22121	{null}	{null}	Dortje	apl. Prof. Dr.	Golldack-Brockhausen	0	Mon Oct 9 21:19:24 2017	Frau
34315	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Gernot	apl. Prof. Dr.	Horstmann	0	Mon Oct 9 21:19:24 2017	Herr
147567	https://ekvv.uni-bielefeld....	https://ekvv.uni-bielefel...	Jörn	apl. Prof. Dr.	Kalinowski	0	Mon Oct 9 21:19:24 2017	Herr
7702352	{null}	{null}	Barbara	apl. Prof. Dr.	Kaltschmidt	0	Mon Oct 9 21:19:24 2017	Frau

Structure of the personal_archive in MongoDB

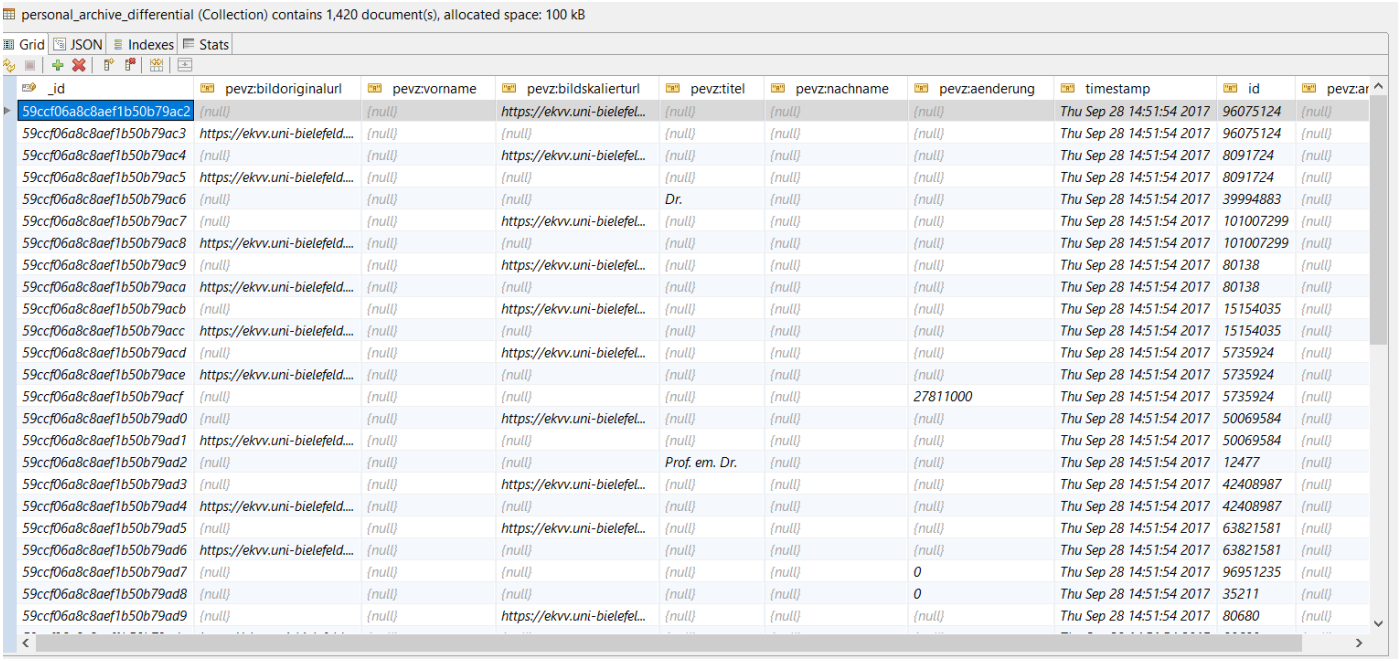


personal_archive (Collection) contains 5,868 document(s), allocated space: 744 kB

Grid | JSON | Indexes | Stats

_id	pevz:bildoriginalurl	pevz:bildskalierturl	pevz:vorname	pevz:titel	pevz:nachname	pevz:aenderung	timestamp	pevz:anrede
107619333	{null}	{null}	Marc	{null}	Eßer	0	Thu Sep 28 15:09:49 2017	Frau
60158951	{null}	{null}	Melanie	{null}	Eulitz	0	Thu Sep 28 15:09:49 2017	Frau
109490137	{null}	{null}	Annika	{null}	Fischer	0	Thu Sep 28 15:09:49 2017	Frau
108493732	{null}	{null}	Nina	{null}	Flottmann	0	Thu Sep 28 15:09:49 2017	Frau
110728789	{null}	{null}	Matthäus	{null}	Fons	0	Thu Sep 28 15:09:49 2017	Herr
108499456	{null}	{null}	Sebastian	{null}	Fuchs	0	Thu Sep 28 15:09:49 2017	Herr
109978162	{null}	{null}	Hendrik	{null}	Füser	0	Thu Sep 28 15:09:49 2017	Frau
107870581	{null}	{null}	Anna	{null}	Funk	0	Thu Sep 28 15:09:49 2017	Frau
108777220	{null}	{null}	Julian	{null}	Gärtner	0	Thu Sep 28 15:09:49 2017	Herr
109694944	{null}	{null}	Lars	Dr.	Gertenbach	0	Thu Sep 28 15:09:49 2017	Herr
111726455	{null}	{null}	Felix	{null}	Geschwinder	54151000	Thu Sep 28 15:09:49 2017	Herr
108938674	{null}	{null}	Leonard	{null}	Gödde	0	Thu Sep 28 15:09:49 2017	Herr
22672373	{null}	{null}	Stephanie	{null}	Görts	0	Thu Sep 28 15:09:49 2017	Frau
109089652	{null}	{null}	Manuel	{null}	Göz	0	Thu Sep 28 15:09:49 2017	Herr
110834774	{null}	{null}	Felix	{null}	Gorny	0	Thu Sep 28 15:09:49 2017	Herr
110290378	{null}	{null}	Hülya	{null}	Grams	0	Thu Sep 28 15:09:49 2017	Frau
107697975	{null}	{null}	Christiana	{null}	Grewe	0	Mon Oct 9 21:19:24 2017	Frau
111449454	{null}	{null}	Thomas	{null}	Grieger	52778000	Thu Sep 28 15:09:49 2017	Herr
106162688	{null}	{null}	Kristin	{null}	Haake	0	Thu Sep 28 15:09:49 2017	Frau
110088520	{null}	{null}	Florian	{null}	Hartmann	0	Thu Sep 28 15:09:49 2017	Herr
108058903	{null}	{null}	Nico	{null}	Hartmann	0	Thu Sep 28 15:09:49 2017	Herr
33886229	{null}	{null}	Tobias	{null}	Henke	0	Thu Sep 28 15:09:49 2017	Herr
109611662	{null}	{null}	Anne-Kristin	{null}	Herling	0	Thu Sep 28 15:09:49 2017	Frau
106484011	{null}	{null}	Martina	Prof. Dr.	Hofmanová	0	Thu Sep 28 15:09:49 2017	Frau
110256934	{null}	{null}	Christopher	{null}	Hohensee	0	Thu Sep 28 15:09:49 2017	Herr

Structure of the personal_archive_differential in MongoDB


personal_archive_differential (Collection) contains 1,420 document(s), allocated space: 100 kB

Grid | JSON | Indexes | Stats

_id	pevz:bildoriginalurl	pevz:vorname	pevz:bildskalierturl	pevz:titel	pevz:nachname	pevz:aenderung	timestamp	id	pevz:ar
59ccf06a8c8aef1b50b79ac2	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	96075124	{null}
59ccf06a8c8aef1b50b79ac3	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	96075124	{null}
59ccf06a8c8aef1b50b79ac4	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	8091724	{null}
59ccf06a8c8aef1b50b79ac5	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	8091724	{null}
59ccf06a8c8aef1b50b79ac6	{null}	{null}	{null}	Dr.	{null}	{null}	Thu Sep 28 14:51:54 2017	39994883	{null}
59ccf06a8c8aef1b50b79ac7	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	101007299	{null}
59ccf06a8c8aef1b50b79ac8	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	101007299	{null}
59ccf06a8c8aef1b50b79ac9	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	80138	{null}
59ccf06a8c8aef1b50b79aca	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	80138	{null}
59ccf06a8c8aef1b50b79acb	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	15154035	{null}
59ccf06a8c8aef1b50b79acc	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	15154035	{null}
59ccf06a8c8aef1b50b79acd	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	5735924	{null}
59ccf06a8c8aef1b50b79ace	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	5735924	{null}
59ccf06a8c8aef1b50b79acf	{null}	{null}	{null}	{null}	{null}	27811000	Thu Sep 28 14:51:54 2017	5735924	{null}
59ccf06a8c8aef1b50b79ad0	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	50069584	{null}
59ccf06a8c8aef1b50b79ad1	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	50069584	{null}
59ccf06a8c8aef1b50b79ad2	{null}	{null}	{null}	Prof. em. Dr.	{null}	{null}	Thu Sep 28 14:51:54 2017	12477	{null}
59ccf06a8c8aef1b50b79ad3	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	42408987	{null}
59ccf06a8c8aef1b50b79ad4	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	42408987	{null}
59ccf06a8c8aef1b50b79ad5	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	63821581	{null}
59ccf06a8c8aef1b50b79ad6	https://ekvv.uni-bielefeld....	{null}	{null}	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	63821581	{null}
59ccf06a8c8aef1b50b79ad7	{null}	{null}	{null}	{null}	{null}	0	Thu Sep 28 14:51:54 2017	96951235	{null}
59ccf06a8c8aef1b50b79ad8	{null}	{null}	{null}	{null}	{null}	0	Thu Sep 28 14:51:54 2017	35211	{null}
59ccf06a8c8aef1b50b79ad9	{null}	{null}	https://ekvv.uni-bielefel...	{null}	{null}	{null}	Thu Sep 28 14:51:54 2017	80680	{null}

Tracking of data:

The system was designed keeping in mind that data which have been changed, are recently added and deleted can be tracked on both record and field level.

New Row Added: Taking the Rows from Personal Table as keys and checking for similar row in personal_archive table. For keys which do not have matches are new.

Rows Deleted: Taking the rows from personal_archive table as keys and checking for similar row in personal table. For keys which do not match are the ones deleted.

Field value updated or added: Taking the Rows from Personal Table as keys and checking for the similar id in personal_archive_differential reveals the number of times a record is changed along with the timestamp of changes.

Field value deleted: Taking the Rows from Personal Table as keys and checking for values with “null” value reveals the field values that are deleted.

The data available were used to find answers to some important queries. Some of them are:

- Total new values added to database.
- Total values archived in database.
- Field which have been changed most number of times with count.
- Id which have been changed most number of times with count.
- Total number of people in CITEC.
- Total number of doctorates in CITEC.
- Total number of recently awarded doctorates in CITEC.
- No. of Person and the list of names who joined Bielefeld University after a certain date.
- List of persons who have left Bielefeld University.

Code snapshot for MongoDB queries:

```
personal_archive_find = db.personal_archive.find({})
counter = personal_archive_find.count()
for doc in personal_archive_find:
    x=db.personal.find({"_id":{ "$eq" :doc.get("_id")}})
    for z in x:
        if z:
            counter = counter - 1

print("\nTotal new values added to database: " + str(counter))

##difference data
f = db.personal_archive_differential.find({})
tag = ""
max = 0
for a in f:
    a_keys = a.keys()
    for key in a_keys:
        if key != "timestamp" and key != "_id" and key != "id" and key != "pevz:aenderung":
            count = db.personal_archive_differential.count({key: {'$exists': 'true'}})
            if count > max:
                max = count
                tag = key

print("\nThe column " + str(tag) + " has been changed " + str(max) + " times.")


f = db.personal_archive_differential.find({})
val = ""
max = 0
for a in f:
    value = a.get("id")
    count = db.personal_archive_differential.count({"id": { "$eq": value }})
    if count > max:
        max = count
        val = value

print("\nThe id " + str(val) + " has been changed " + str(max) + " times.")
```

Results:

```
Total new values added to database: 30

The column pevz:bildskalierturl has been changed 366 times.

The id 5735924 has been changed 18 times.

Total number of people in CITEC: 5891

Total number of doctorates in CITEC: 1499

Total number of recently awarded doctorates in CITEC: 137

No. of Person who joined Bielefeld University after: 2017-10-07 is: 18

List of persons: ['Tanja Adam-Ashuri', 'Linda Berker', 'Eike Friederike Eifler', 'Ludwig Elsner',
'Christiana Grewe', 'Claudia Mertens', 'Sandra Neufeld', 'Agnes Piekacz', 'Florian Polkowski',
'Daniel Pultermann', 'Heike Quentmeier', 'Karin Raker', 'Kerstin Rehr', 'Andreas Rempel', 'Karl
Rohlfs', 'Paulo Astor Soethe', 'Wilena Telman', 'Lara Thomas']

Total values archived in database: 56

List of persons: ['Dato Abashidze', 'Karim Abdelhak', 'Negera Abdissa Ayana', 'Jürgen Abel',
'Marcus Abel', 'Thomas Abel', 'Christian Abeling', 'He
rbert Abels', 'Werner Abelshauser', 'Ferhat Acar', 'Jascha Achenbach', 'Kathrin Ackermann',
'Sabine Adam', 'Timo Adam', 'Tanja Ada
m-Ashuri', 'Anita Adamczyk', 'Sarah Adameh', 'Andrea Adams', 'Julian Adams', 'Michael Adams',
'Sebastian Adloff', 'Sylvia Agbih', 'Elena Aguiar', 'N.N. AG 4', 'Ailar Ahangri', 'Gerhard
Ahlers', 'Carolin Ahlert', 'Shabnam Ahmadzai', 'Mokhtar Ahmed', 'Jutta Ahrendt', 'Tobias
```

```
Ahsendorf', 'Zeynep Akbayin', 'Firat Akbulut', 'Nurcan Akbulut', 'Gernot Akemann', 'Fatma Akkaya-
Willis', 'Mustafa A
ksakal', 'Serdal Alabas', 'Ahmad Al Ajlan', 'Stefan Albaum', 'Imke Albers', 'Andreas
Albersmeier', 'Gleb Albert', 'Mathias Albert'
, 'Lothar Albertin', 'Sergio Albeverio', 'Günter Albrecht', 'Hans-Jörg Albrecht', 'Melanie
Albrecht', 'Oliver Albrecht', 'Petra Al
brecht', 'Bisan Al Bunni', 'Eva Kristina Albus', 'Stefanie Albus', 'Annette von Alemann',
'Alexander Alempic']
```

4. Repository

https://github.com/svyneet/version_control_of_institutiondata_with_mongodb

5. Conclusion

To put it in a nutshell the project has activities like selection of type of archiving and dataset to be stored so that the tracking of data on different levels can be achieved. It was an interesting experience to use the data and find out realistic questions from them. The queries were selected on the basis of importance of answers they can provide. The dataset can be explored and extended to answer more different questions.