



Data safety and standardization

Lecture overview

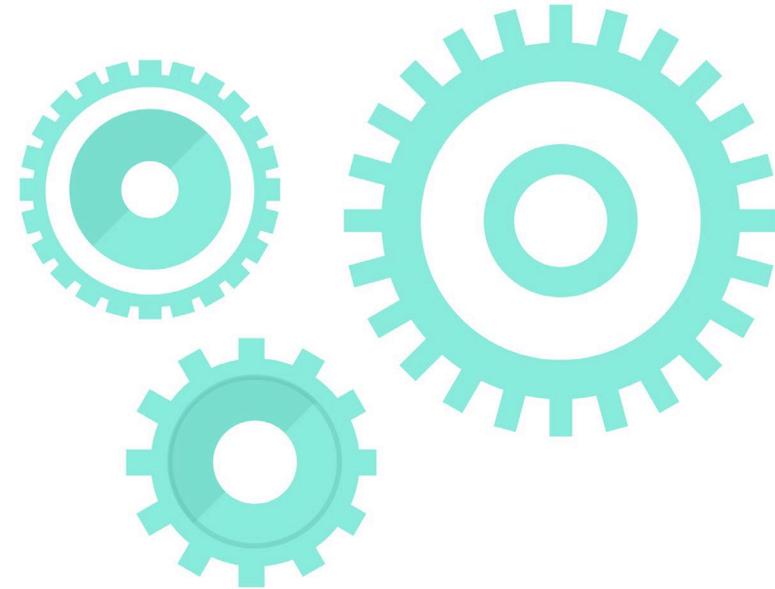
- Importance of concepts in lecture
- Data storage
 - methods, formats, databases
- Data interfaces
 - Web, software, APIs
- Data standardization
- Data security
- Data management plants (DMPs)

Importance of Data safety

- Data can be sensitive
- Data can be exploited
- Privacy issues

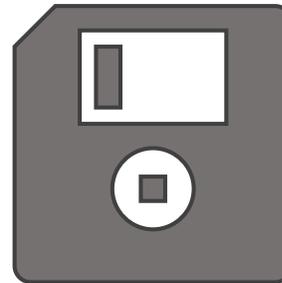
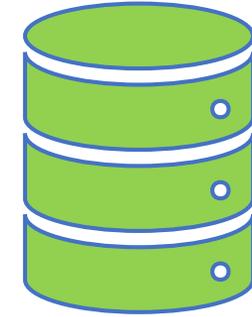
Importance standardization

- Makes data comparison easier
- Lessons effort in data collection
- Can automate systems



Data storage methods

- Data is stored in a number of ways
- There are advantages and disadvantages
 - Ease of access
 - Storing efficiency
 - Access controls



Data storage formats - Text

- Text based formats
 - CSV, JSON, XML +++
- Stores data in ASCII or UTF/Unicode text
- Easy to use
- Disadvantages of text storage formats
 - No security features
 - Inefficient in terms of speed and space for large data

```
- <EmployeeData>
  - <employee id="34594">
    <firstName>Heather</firstName>
    <lastName>Banks</lastName>
    <hireDate>1/19/1998</hireDate>
    <deptCode>BB001</deptCode>
    <salary>72000</salary>
  </employee>
  - <employee id="34593">
    <firstName>Tina</fir
    <lastName>Young</
    <hireDate>4/1/2010
    <deptCode>BB001</
    <salary>65000</sal:
  </employee>
</EmployeeData>
```

```
File Edit Format View Help
Index,Make,Model,Year
0,Ford,Crown Victoria,1993
1,Ford,Taurus,1980
2,Tesla,Model 3,2017
3,Tesla,Model S,2018
4,Tesla,Model X,2019
5,Tesla,Roadster,2019
6,Toyota,Camry,1994
7,Toyota,Celica,1990
8,Toyota,Tundra,2001
9,Honda,Accord,2005
10,Kia,Rio,2011
11,Hyundai,Elantra,2006
12,Nissan,Pathfinder,1997
```

```
{
  "array": [
    1,
    2,
    3,
    4
  ],
  "boolean": true,
  "color": "#82b92c",
  "null": null,
  "number": 123,
  "object": {
    "a": "b",
    "c": "d",
    "e": "f"
  },
  "string": "Hello World"
}
```

Comma separated values (CSV)

- Stores table based data where each column is typically separated by a comma “,”.
- In locales where comma is used as a decimal separator the separator is often a semi-colon “;” instead.
- Spreadsheet software can open these files and represent them as tables.
- File types/endings are either .csv or .txt.
- Can be challenging to store more complicated datatypes with relations

```
File Edit Format View Help
Index,Make,Model,Year
0,Ford,Crown Victoria,1993
1,Ford,Taurus,1980
2,Tesla,Model 3,2017
3,Tesla,Model S,2018
4,Tesla,Model X,2019
5,Tesla,Roadster,2019
6,Toyota,Camry,1994
7,Toyota,Celica,1990
8,Toyota,Tundra,2001
9,Honda,Accord,2005
10,Kia,Rio,2011
11,Hyundai,Elantra,2006
12,Nissan,Pathfinder,1997
```

Characteristics of CSV data

- Requires little management
- Best for curated data sets
- EU research CSV download
 - <https://data.europa.eu/data/datasets?locale=en&minScoring=0&query=water&page=1&format=CSV>

Java Script Object Notation (JSON)

- Stores structured data using in so called JSON object. Each **object** is encased in curly brackets “{}” and consists of **keys and values** separated by a colon “:”. Each element is separated by a comma “,”.
- Can store more complicated data as the values can be JSON objects themselves.
- Commonly used to transferring data over the internet

```
{"City": "Oslo",  
  "details": {  
    "long": 41.23,  
    "lat": 45.13  
  }  
}
```

eXtensible Markup Language (XML)

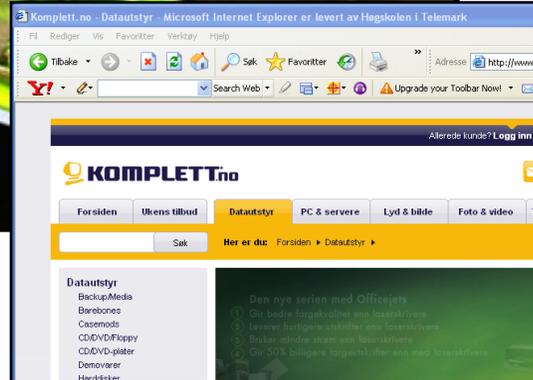
- Similar to JSON, but uses different syntax. Instead of keys, values are encased in a Markup elements called **tag** that contain essentially the **key**.
- Tags are ended by a /
- Each tag can have **attributes** as well
- Commonly used for transferring data on the internet

```
- <EmployeeData>
  - <employee id="34594">
    <firstName>Heather</firstName>
    <lastName>Banks</lastName>
    <hireDate>1/19/1998</hireDate>
    <deptCode>BB001</deptCode>
    <salary>72000</salary>
  </employee>
  - <employee id="34593">
    <firstName>Tina</firstName>
    <lastName>Young</lastName>
    <hireDate>4/1/2010</hireDate>
    <deptCode>BB001</deptCode>
    <salary>65000</salary>
  </employee>
</EmployeeData>
```

Database Management System (DBMS)

- Broad term for any system that enables storage of structured data in an efficient manner.
- There are multiple DBMS classifications
 - Relation database, non-relations database, Document database, network database, graph database and Hierarchical.
- Most common examples are of the relational kind, but there are new emerging databases gaining traction such as noSQL.
- Regardless of the underlying structure we typically communicate with these databases using a Structured Query language (SQL).

There are often databases underneath



YR.no Søkk blant 7 millioner værvarsel i Norge og heile verda: **VÆRSØK**

Skriv stadnamn, f.eks. Stavanger, Røst eller Beijing.

Varsel for Norge

- Østlandet
- Sørlandet
- Vestlandet
- Trøndelag
- Nord-Norge
- Polarområdene
- Skiføre (snøkart) **NY!**
- Snødybder
- Pollenvarsel

OBS-varsel: Vanskelige kjøreforhold i Vest-Finnmark fredag.
28.03.2008 06:50: Fredag kan det bli lokalt vanskelige kjøreforhold på grunn av sterk vind, snøbyger og snøfokk. .

| Andre steder | Fredag | Lørdag | Søndag |
|---------------------------|--------|--------|--------|
| Oslo | 2° | 4° | 5° |
| Bergen | 5° | 4° | 7° |
| Stavanger | 6° | 6° | 7° |

VARMEST-KALDAST-VÅTAST

| | |
|----------------------------------|----------|
| 27. mars kl 7 til 19 | 6,7 °C |
| Dagali | -27,6 °C |
| Bergen/Flesland | 8,0 mm |
| 27. mars kl 19 til 28. mars kl 7 | |
| Obrestad Fyr | 4,6 °C |
| Sihcjavri | -30,1 °C |
| Nedre Vats | 21,9 mm |

Kva er yr.no?



Relational databases

- Relational databases store data in tables with rows and columns. A database can have multiple tables.
- These tables typically have to be defined in advance.
- Each column has a certain data type (whole number, text etc...) defined in advance. This increases efficiency.
- What makes relational databases stand out is the ability to relate data in one table's column to another table's column.
- When making queries on the database, these relations can be used to connect multiple sources of data together.
- As tables have to be defined in advance these types of DBMS are inflexible.

Relations

Relation



Lender table

| Lender_ID | Lender_name | ISBN |
|-----------|-------------|------|
| | | |
| | | |
| | | |

Book table

| ISBN | Book_name | Author |
|------|-----------|--------|
| | | |
| | | |
| | | |

Non relational database (noSQL)

- A DBMS built on using the JSON structure.
- There are no direct relations.
- Flexible as they don't have predefined structures as with relational databases



Benefits of DBMS

- Each data type is defined in a DBMS which speeds up data search and collection for big data sets.
- DBMSs implement effective lookup procedures through their SQL commands. This allows us not to worry about which algorithms to use when accessing the data.

Data Interfaces

- DBMS and file systems are often not connected to directly by users.
- Often an interface sits between the database and the user.
- Interfaces
 - web interface
 - software interface
- Application Programming Interface (APIs)
- Software interfaces usually get data from an online source so the difference isn't that big.

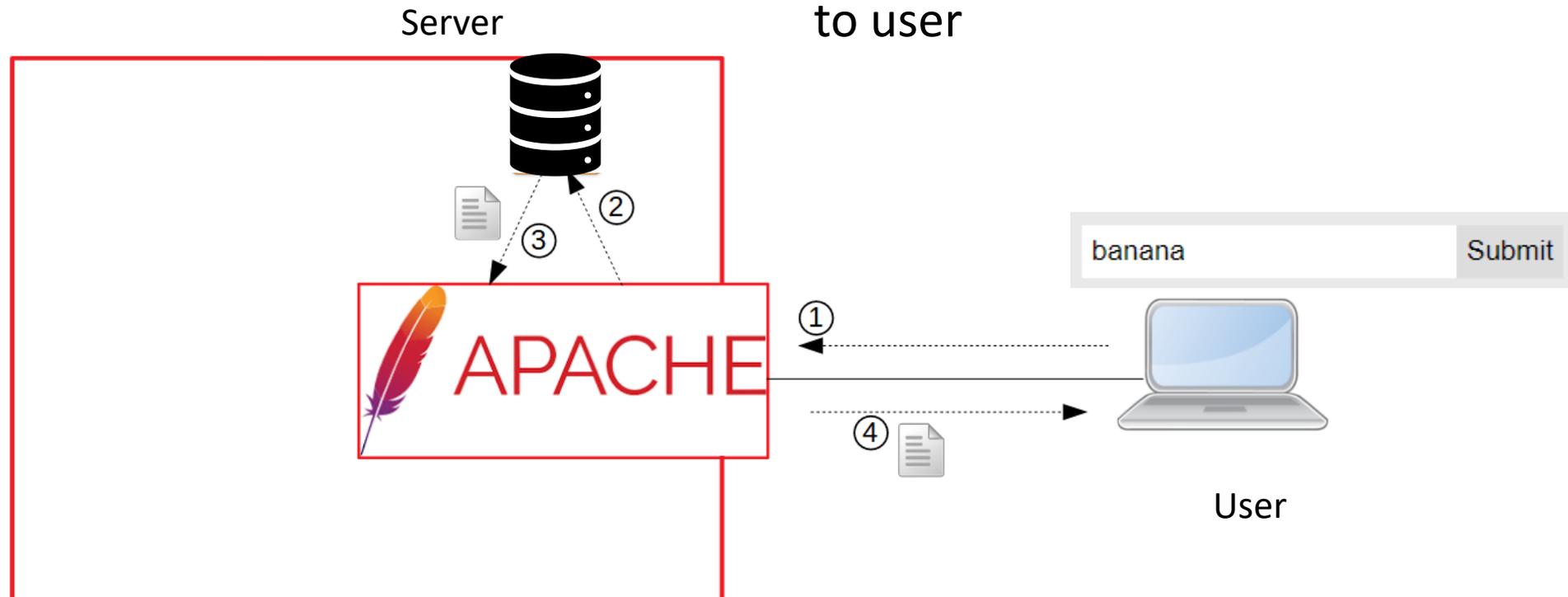
A simple web interface

1 – User requests data

3 – database transfers data to server handler

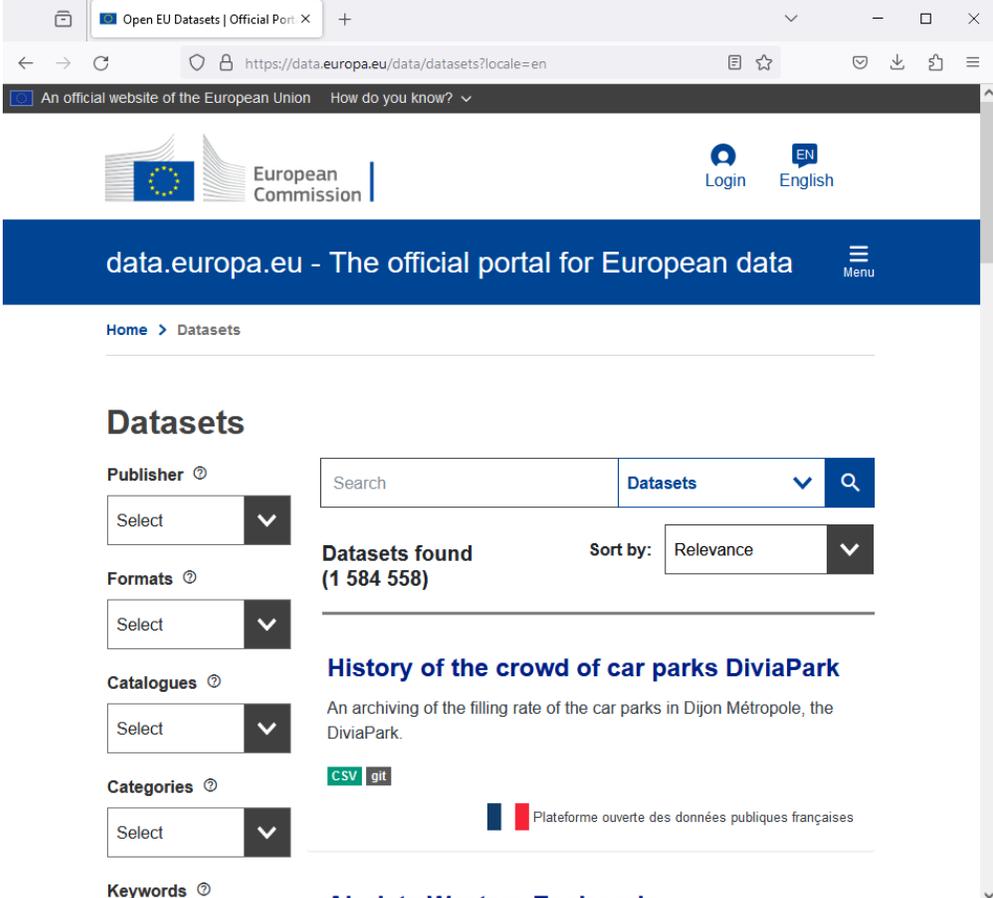
2 – server handler passes request to database

4 – server handler passes on data to user



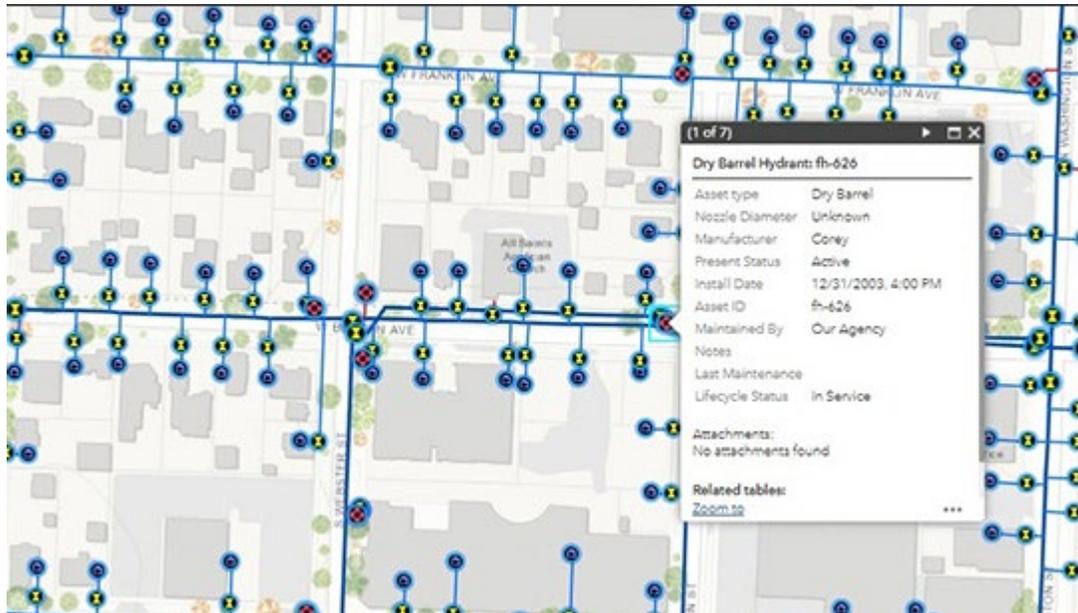
Web interface example

- data.europa.eu
 - Most likely has a combination of databases and files based storage system

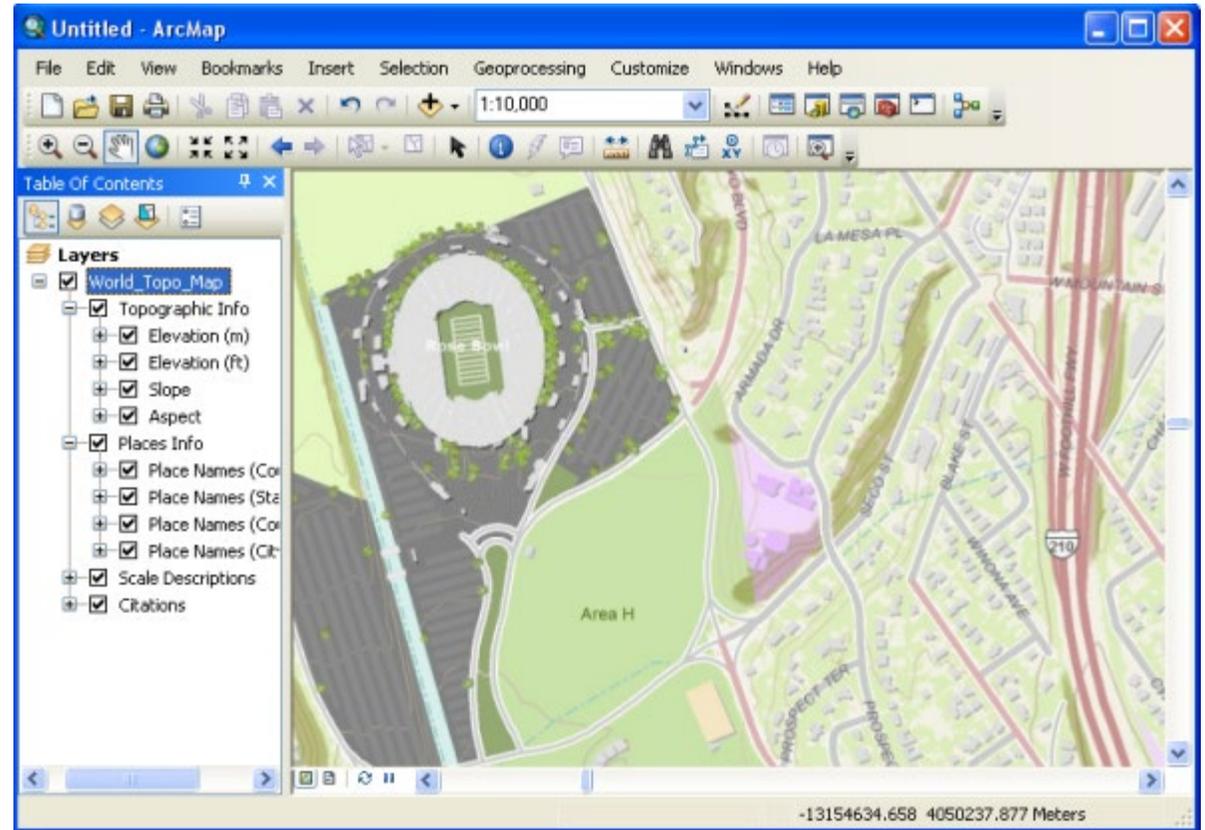


The screenshot shows the data.europa.eu website interface. The browser address bar displays the URL <https://data.europa.eu/data/datasets?locale=en>. The page header includes the European Commission logo, a 'Login' button, and a language selector set to 'English'. Below the header, a blue navigation bar contains the text 'data.europa.eu - The official portal for European data' and a 'Menu' icon. The main content area is titled 'Datasets' and features a search bar with the text 'Search' and a dropdown menu set to 'Datasets'. Below the search bar, there are four filter sections: 'Publisher', 'Formats', 'Catalogues', and 'Categories', each with a 'Select' dropdown menu. To the right of the search bar, the text 'Datasets found (1 584 558)' is displayed, along with a 'Sort by:' dropdown menu set to 'Relevance'. Below the filters, a featured dataset is shown with the title 'History of the crowd of car parks DiviaPark' and a description: 'An archiving of the filling rate of the car parks in Dijon Métropole, the DiviaPark.' The dataset is available in 'CSV' and 'git' formats. At the bottom of the page, there is a logo for 'Plateforme ouverte des données publiques françaises' and a partially visible title 'Air data Western Espenada'.

Software interfaces



esri.com



arcgis.com

Application Programming Interface (APIs)

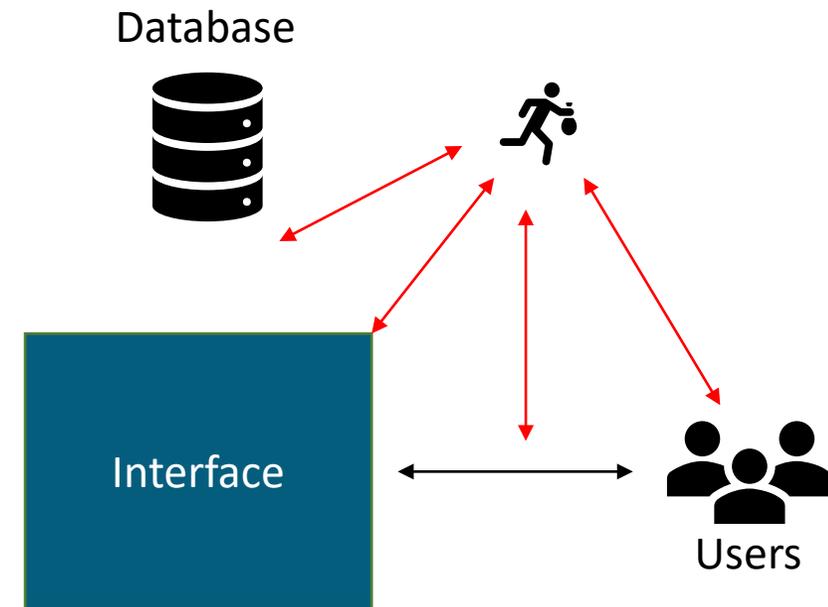
- Sometimes you want a machine to do the handling of the data transfer and processing. This makes web interfaces and software interfaces hard to use.
- For this there is a concept known as API which allows for machine communication of data.
- Examples:
 - <https://openweathermap.org/api>
 - <https://wikis.ec.europa.eu/display/EUROSTATHELP/API+-+Getting+started>

Data standardization

- With so many different data storage options and storage metrics it can be challenging to compare and use data.
- Without interfaces there are knowledge requirements in how to access and process the data
- The FAIR data management principles
- <https://www.go-fair.org/fair-principles/>

Data vulnerability

- Data is vulnerable to attack at several stages.
- Data theft
- Valuable data
 - Security concerns
 - Personal data
- Data corruption
 - Ransomware



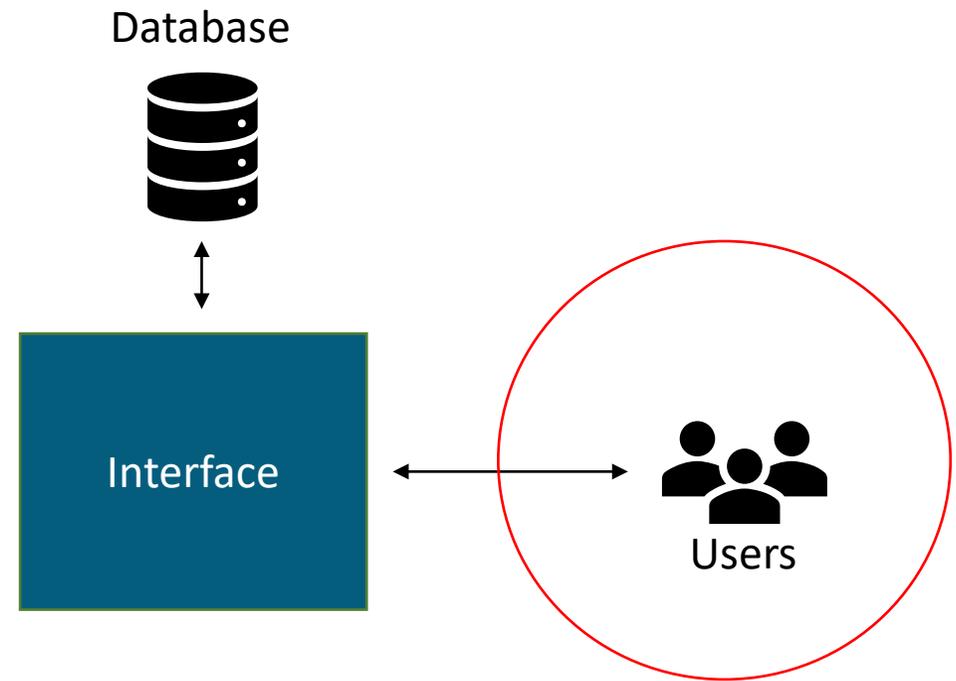
Data security

- Access control/authentication
- Encryption
- Monitoring
- Backup
- Action plans



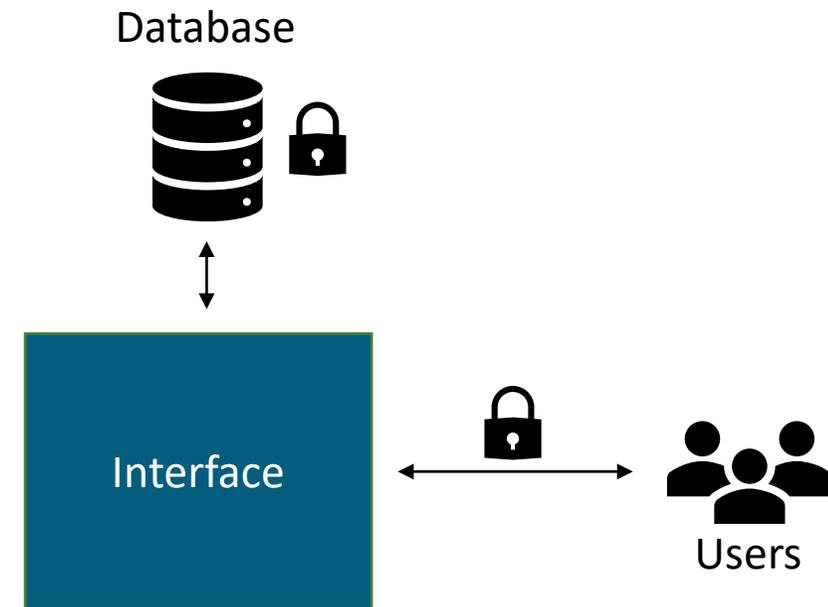
Access control/authentication

- Ensure that only authorized users can access.



Encryption

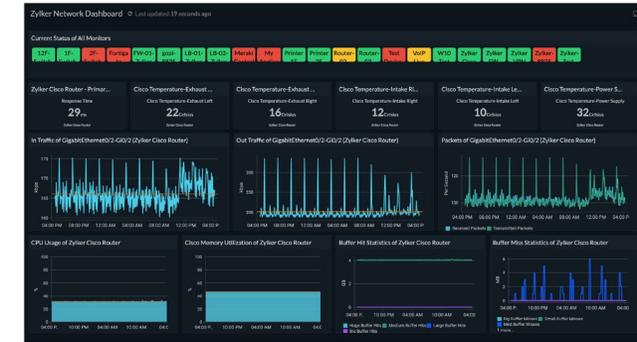
- Prevent data from being read correctly without authorized access.
- Scramble data both in storage and in transit.



Monitoring

- Not all systems are perfect. There will be bugs.
- Important to monitor the system for unusual user patterns to detect intrusions.

Zyker



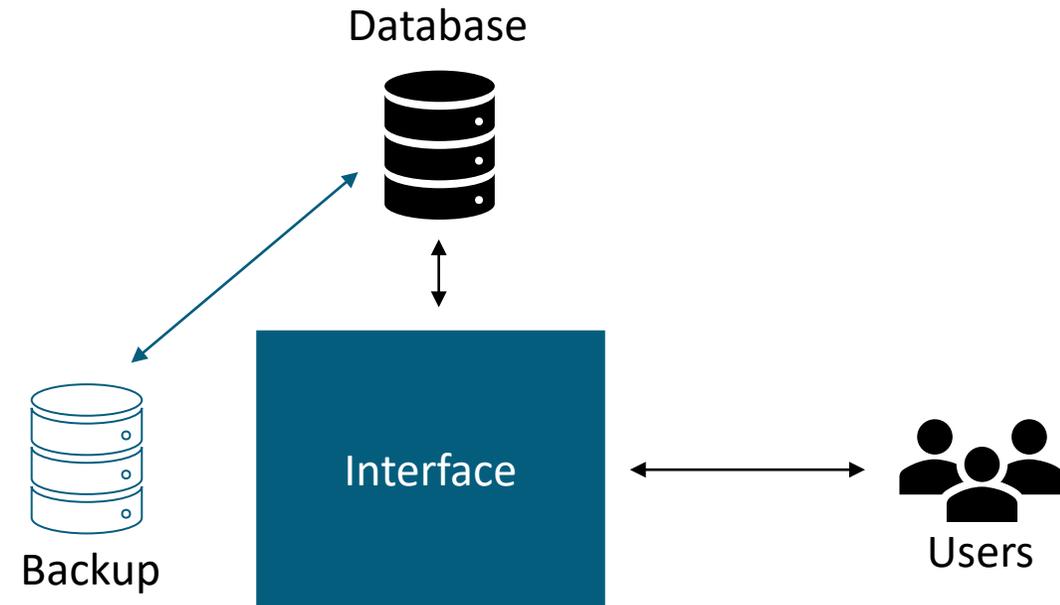
Database



Users

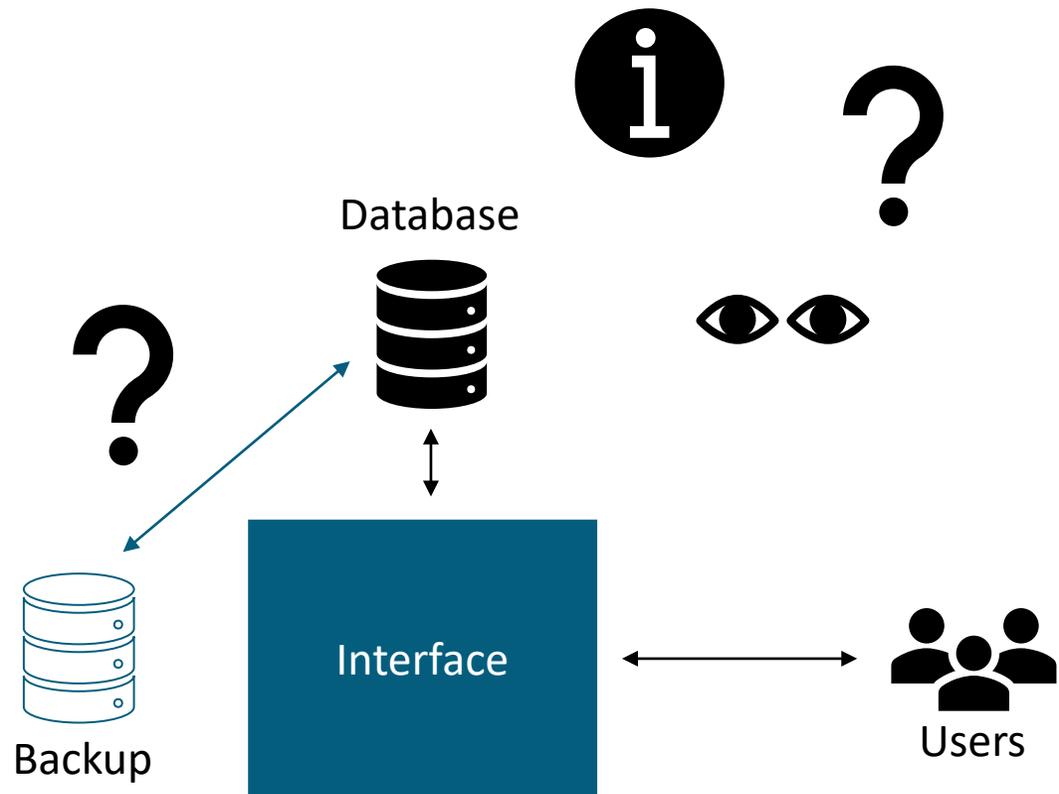
Backup

- Sometimes corruption happens. In those cases it is a great idea to have backups



Action plans

- When an intrusion happens, it is important to have an action plan to follow.
- Data restoration
- Notification
- Access revoke
- Practices review



Data management plans (DMP)

- Projects, research and processes generate lots of data
- We can improve the benefits from these activities by handling the data in a structured way
- It is a good idea and often required to create data management plans (DMP)

Points for DMP

- Who is responsible for the data in its different lifecycle steps
- How you ensure the data is organized and documented
- The amount and type of data generated
- Security of the data in its lifetime
- Metadata and data interoperability
- Ethical considerations
- Long term storage and availability

Considerations for the water sector

- Ethical/data sharing
 - Sharing data could expose private information
 - Example – Drug use estimate from sewage: https://www.emcdda.europa.eu/publications/html/pods/waste-water-analysis_en
- Meta data
 - Correctly labelling data sources can allow for the data to be used in larger studies after project completion.

DMP resources

- European Commission

- https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

- Sikt

- <https://sikt.no/en/data-management-plan>

- Word template for DMP (EU)

- https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

- DMP example

- <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bfe68de0&appld=PPGMS>

- Data repositories for storage

- <https://www.re3data.org/> , <https://fairsharing.org/>