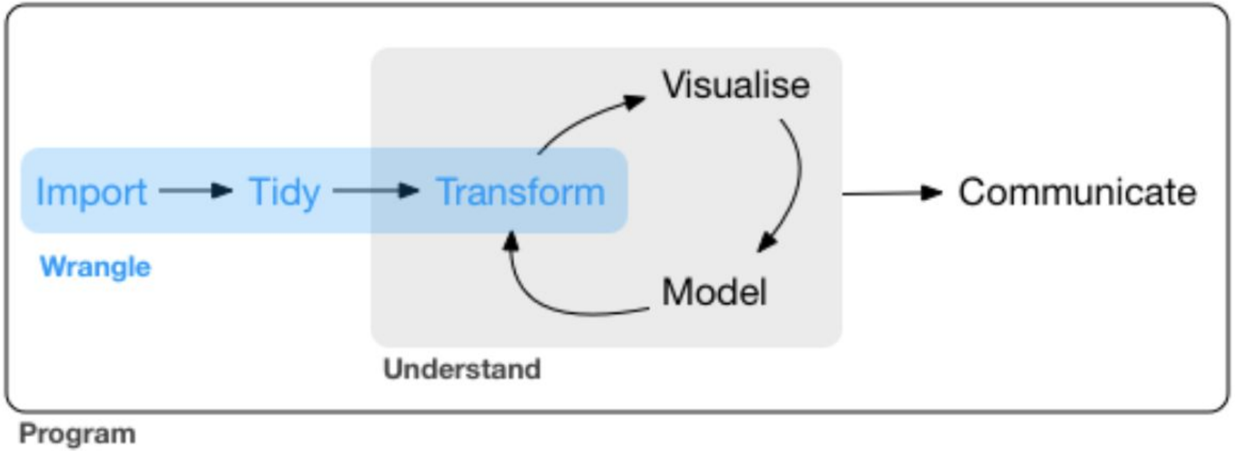
1.2 Import

Applied Data Analysis (ADA)

Oxford DH Summer School - 2023



https://r4ds.had.co.nz/introduction.html

Data formats

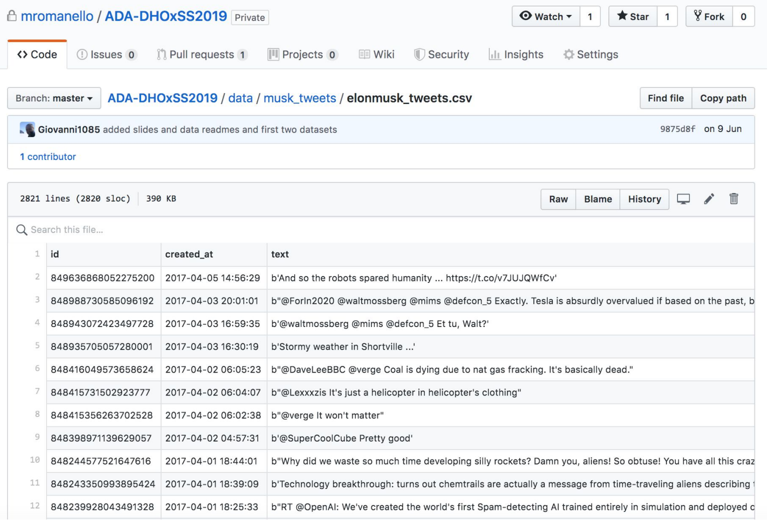
In any data analysis project, a substantial chunk of time goes into preparing your data for analysis. This includes reading (legacy) data formats, storing data to intermediates files, saving data to a database system for longer-term storage.

We will focus on the following **key formats**:

- CSV
- XML
- JSON

CSV

- *CSV*: comma separated values
- *TSV*: tab separated values
- The *lingua franca* of tabular data
- First row: column headers



Giovanni1085 added slides and data readmes and first two datasets 9875d8f on 9 Jun

1 contributor

2821 lines (2820 sloc) 390 KB

id	created_at	text
849636868052275200	2017-04-05 14:56:29	b'And so the robots spared humanity ... https://t.co/v7JUJQWfCv'
848988730585096192	2017-04-03 20:01:01	b"@ForIn2020 @waltmossberg @mims @defcon_5 Exactly. Tesla is absurdly overvalued if based on the past, b
848943072423497728	2017-04-03 16:59:35	b'@waltmossberg @mims @defcon_5 Et tu, Walt?'
848935705057280001	2017-04-03 16:30:19	b'Stormy weather in Shortville ...'
848416049573658624	2017-04-02 06:05:23	b"@DaveLeeBBC @verge Coal is dying due to nat gas fracking. It's basically dead."
848415731502923777	2017-04-02 06:04:07	b"@Lexxxzis It's just a helicopter in helicopter's clothing"
848415356263702528	2017-04-02 06:02:38	b"@verge It won't matter"
848398971139629057	2017-04-02 04:57:31	b'@SuperCoolCube Pretty good'
848244577521647616	2017-04-01 18:44:01	b"Why did we waste so much time developing silly rockets? Damn you, aliens! So obtuse! You have all this craz
848243350993895424	2017-04-01 18:39:09	b'Technology breakthrough: turns out chemtrails are actually a message from time-traveling aliens describing
848239928043491328	2017-04-01 18:25:33	b"RT @OpenAI: We've created the world's first Spam-detecting AI trained entirely in simulation and deployed c

Some **drawbacks**:

- Characters in noisy text columns may interfere with separators (use TSV)
- Data type information not stored in the file (unlike e.g. Parquet)

XML

e**X**tensible **M**arkup **L**anguage

Descriptive markup: focus on **content** (data) rather than its **presentation**

Extensible: there is no pre-defined tagset

XML documents:

- must be well-formed (document has one root; all elements closing tag, etc.)
- must validate against a schema (structure first!)

XML

Elements

(1) `<l>The sun does arise, </l>`

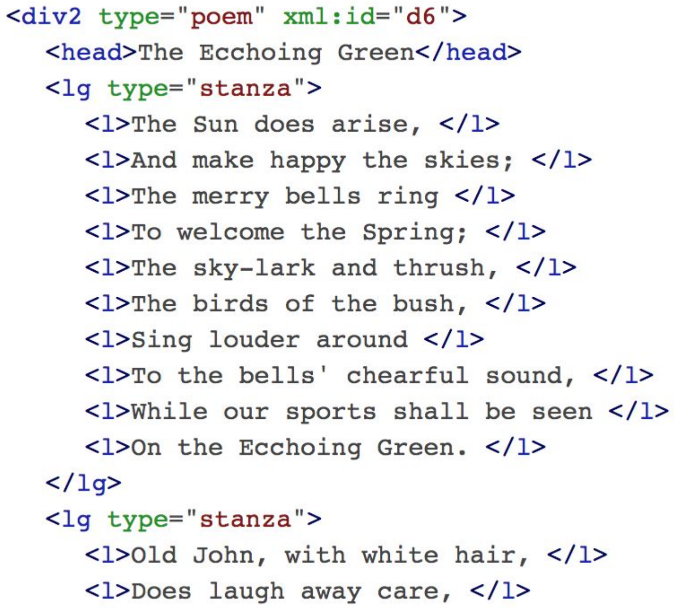
(2) `<br />`

(3)
```
<lg>
    <l>The sun does arise, </l>
  </lg>
```

Attributes

- Element
- Attribute
- attribute value

```
<lg type="stanza">
```



```
<div2 type="poem" xml:id="d6">
  <head>The Ecchoing Green</head>
  <lg type="stanza">
    <l>The Sun does arise, </l>
    <l>And make happy the skies; </l>
    <l>The merry bells ring </l>
    <l>To welcome the Spring; </l>
    <l>The sky-lark and thrush, </l>
    <l>The birds of the bush, </l>
    <l>Sing louder around </l>
    <l>To the bells' chearful sound, </l>
    <l>While our sports shall be seen </l>
    <l>On the Ecchoing Green. </l>
  </lg>
  <lg type="stanza">
    <l>Old John, with white hair, </l>
    <l>Does laugh away care, </l>
```

Mixing vocabularies:

Declaration of namespaces in the root element

```
<root xmlns:tei="http://www.tei-c.org/ns/1.0">
<tei:lg>
```

XML

Technologies:

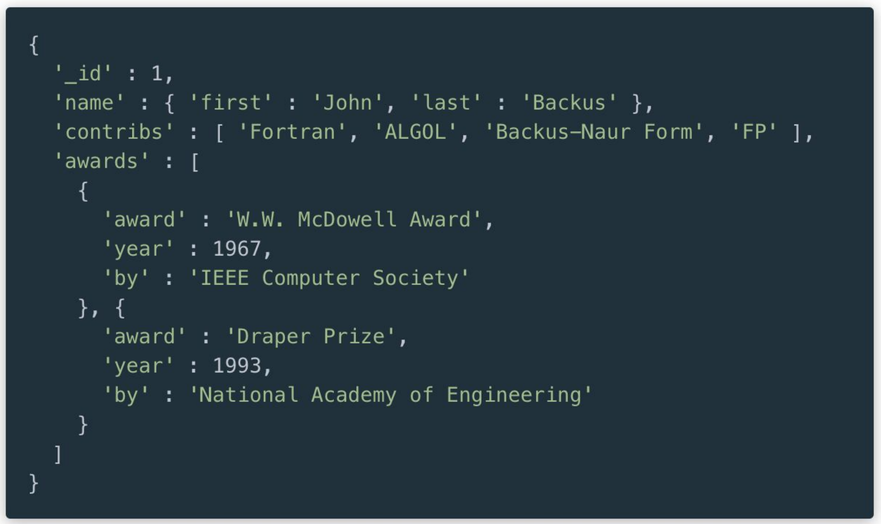
- *Presentation* → CSS
- *Transformation* → XSLT
- *Navigation* → XPath
- *Query* → XQuery

XML in the wild:

- TEI (Text Encoding Initiative) XML
- MARCXML (library catalogue data)
- RDF XML
- ALTO XML (OCR data)
- GraphML
- ...

JSON

Java**S**cript **O**bject **N**otation



```
{
  '_id' : 1,
  'name' : { 'first' : 'John', 'last' : 'Backus' },
  'contribs' : [ 'Fortran', 'ALGOL', 'Backus-Naur Form', 'FP' ],
  'awards' : [
    {
      'award' : 'W.W. McDowell Award',
      'year' : 1967,
      'by' : 'IEEE Computer Society'
    }, {
      'award' : 'Draper Prize',
      'year' : 1993,
      'by' : 'National Academy of Engineering'
    }
  ]
}
```

https://www.mongodb.com/json-and-bson

Data types:

- Number
- String: any sequence of zero or more Unicode characters
- Boolean: true | false
- Array (list in Python)
- Object: unordered collection of name-value pairs {“title”: “ADA DHOxSS”}

JSON

- Need for structure: JSON schema
- Databases that *speak* JSON: ElasticSearch, MongoDB, etc.
- JSON-LD: Web-friendly format to encoded RDF data

Working with data formats

(In ADA most of the times we don’t get to choose the format of data we work with.)

But if you are the one to choose, consider:

- Target use (analysis, online presentation, etc.)
- Target community (wider public, scholars, data scientists)
- Multiple formats for multiple usage scenarios (internal usage, data publication, web application, etc.)