Data Quality and Data Curation – a personal view

Kevin Ashley
Director, Digital Curation Centre
www.dcc.ac.uk
Kevin.ashley@ed.ac.uk

The DCC is supported by Jisc

DCC because good research needs good data

(An aside – terminology)

Data repository

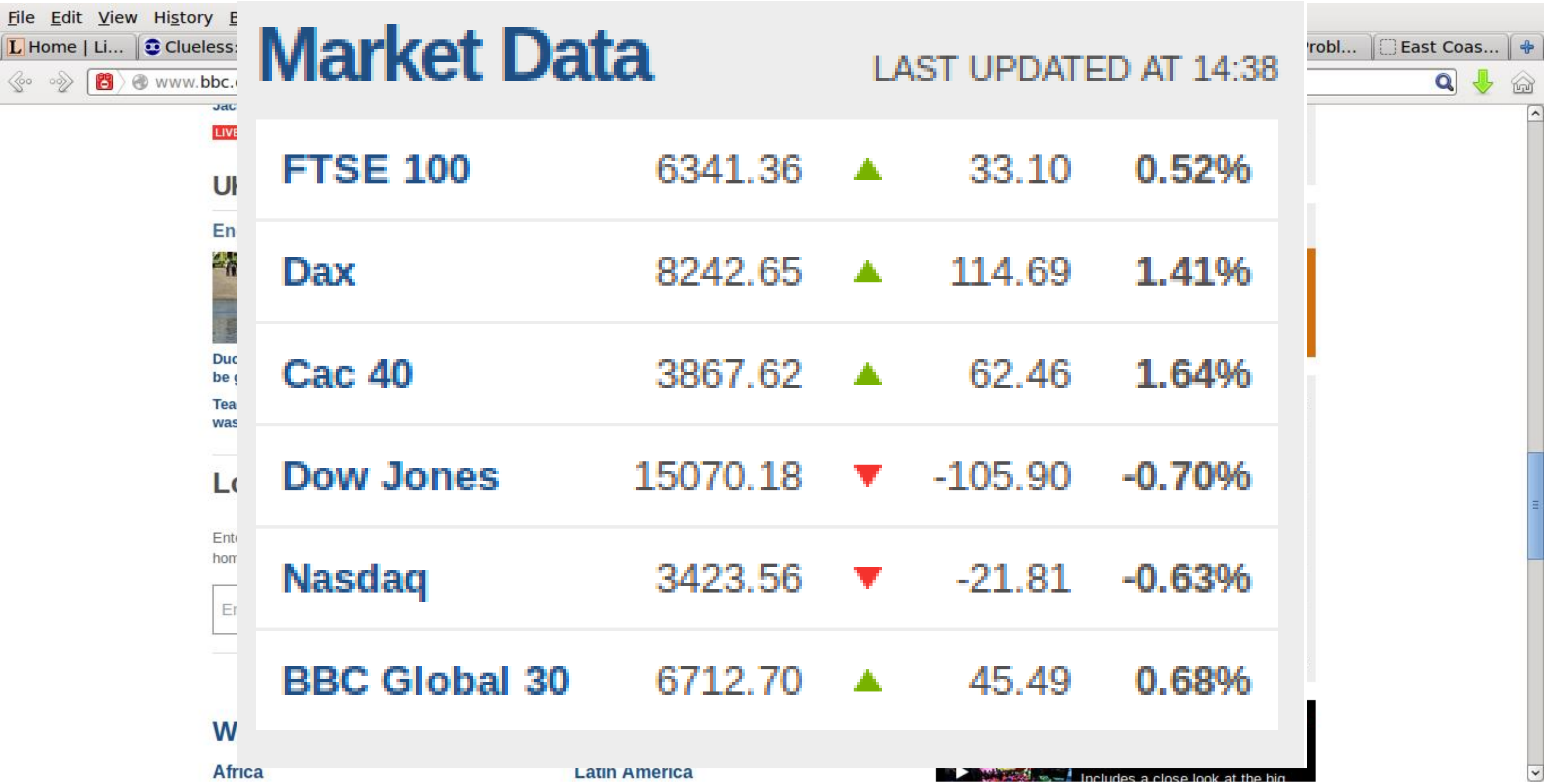
Data archive

Data library

Data bank

Data center

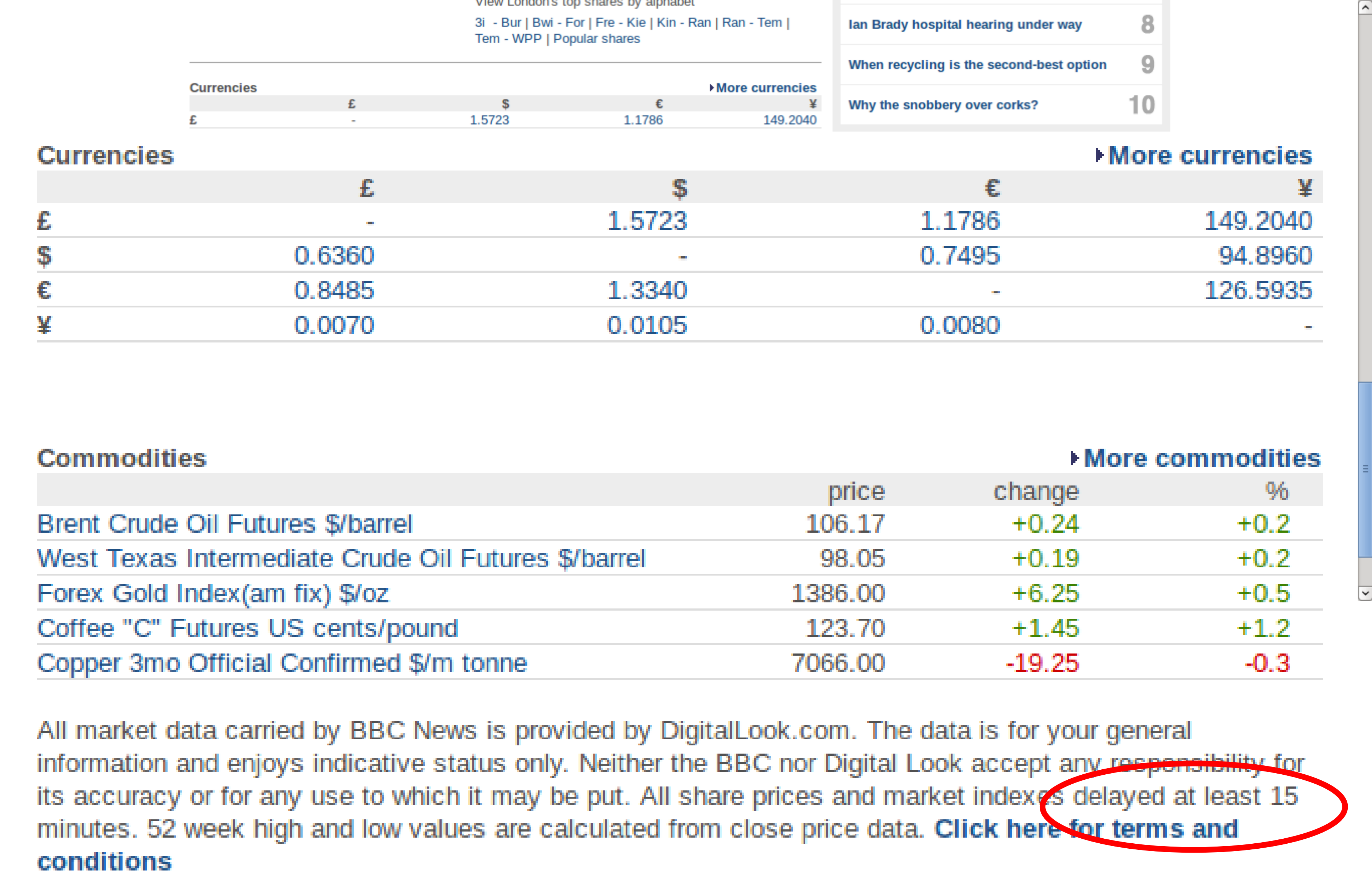
I will use these terms interchangeably



Market Data

LAST UPDATED AT 14:38

FTSE 100	6341.36	▲	33.10	0.52%
Dax	8242.65	▲	114.69	1.41%
Cac 40	3867.62	▲	62.46	1.64%
Dow Jones	15070.18	▼	-105.90	-0.70%
Nasdaq	3423.56	▼	-21.81	-0.63%
BBC Global 30	6712.70	▲	45.49	0.68%



View London's top shares by alphabet

3i - Bur | Bwi - For | Fre - Kie | Kin - Ran | Ran - Tem | Tem - WPP | Popular shares

Ian Brady hospital hearing under way	8
When recycling is the second-best option	9
Why the snobbery over corks?	10

Currencies ▸More currencies

	£	$	€	¥
£	-	1.5723	1.1786	149.2040

Currencies

▸More currencies

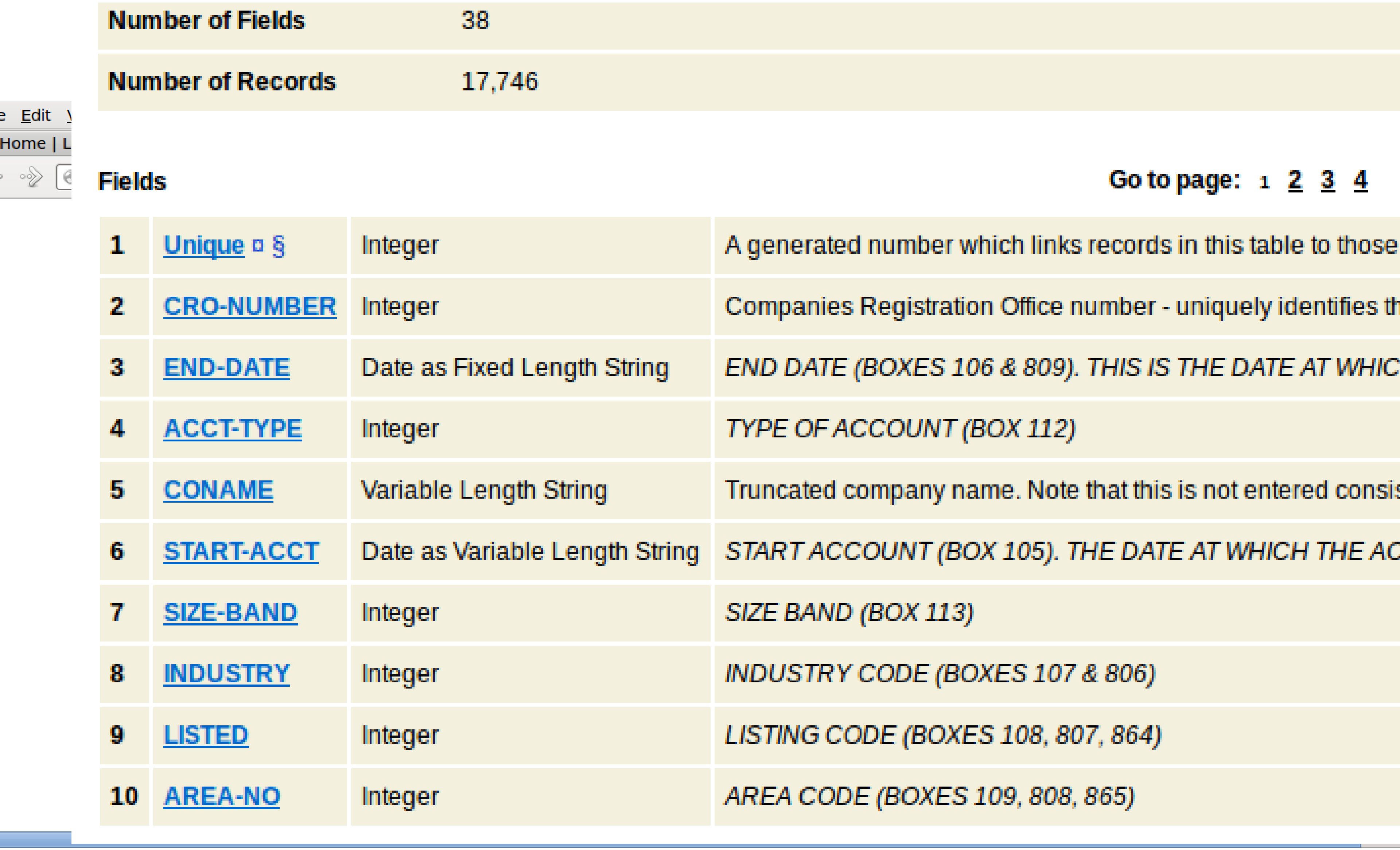
	£	$	€	¥
£	-	1.5723	1.1786	149.2040
$	0.6360	-	0.7495	94.8960
€	0.8485	1.3340	-	126.5935
¥	0.0070	0.0105	0.0080	-

Commodities

▸More commodities

	price	change	%
Brent Crude Oil Futures $/barrel	106.17	+0.24	+0.2
West Texas Intermediate Crude Oil Futures $/barrel	98.05	+0.19	+0.2
Forex Gold Index(am fix) $/oz	1386.00	+6.25	+0.5
Coffee "C" Futures US cents/pound	123.70	+1.45	+1.2
Copper 3mo Official Confirmed $/m tonne	7066.00	-19.25	-0.3

All market data carried by BBC News is provided by DigitalLook.com. The data is for your general information and enjoys indicative status only. Neither the BBC nor Digital Look accept any responsibility for its accuracy or for any use to which it may be put. All share prices and market indexes delayed at least 15 minutes. 52 week high and low values are calculated from close price data. **Click here for terms and conditions**

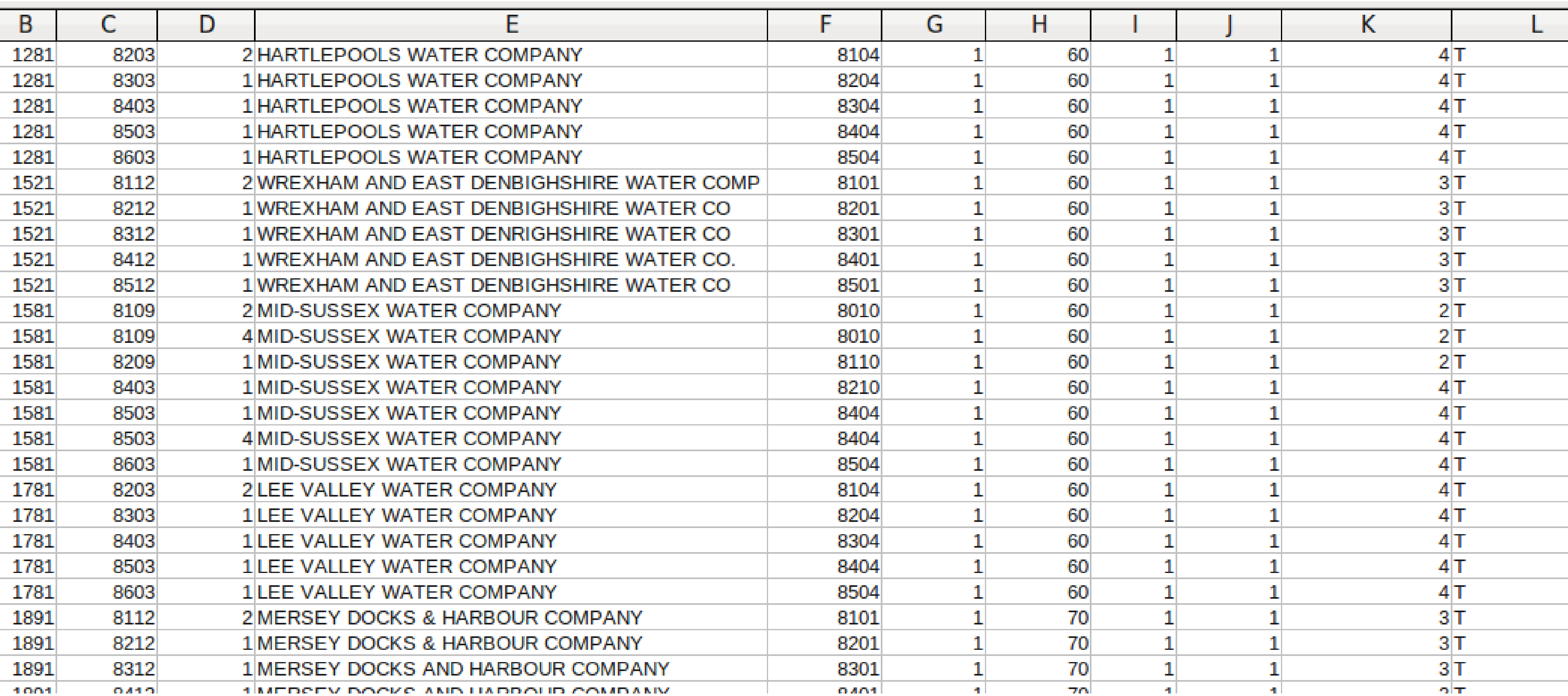


Number of Fields	38
Number of Records	17,746

Fields

Go to page: 1 **2** **3** **4**

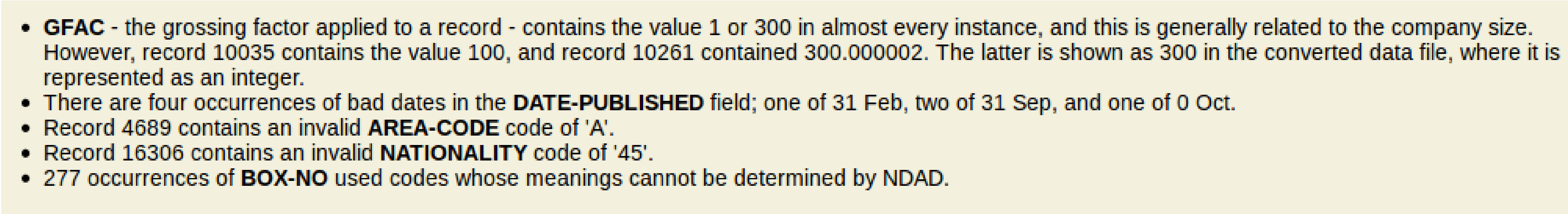
1	Unique ¤ §	Integer	A generated number which links records in this table to those
2	CRO-NUMBER	Integer	Companies Registration Office number - uniquely identifies th
3	END-DATE	Date as Fixed Length String	*END DATE (BOXES 106 & 809). THIS IS THE DATE AT WHIC*
4	ACCT-TYPE	Integer	*TYPE OF ACCOUNT (BOX 112)*
5	CONAME	Variable Length String	Truncated company name. Note that this is not entered consis
6	START-ACCT	Date as Variable Length String	*START ACCOUNT (BOX 105). THE DATE AT WHICH THE AC*
7	SIZE-BAND	Integer	*SIZE BAND (BOX 113)*
8	INDUSTRY	Integer	*INDUSTRY CODE (BOXES 107 & 806)*
9	LISTED	Integer	*LISTING CODE (BOXES 108, 807, 864)*
10	AREA-NO	Integer	*AREA CODE (BOXES 109, 808, 865)*



B	C	D	E	F	G	H	I	J	K	L
1281	8203	2	HARTLEPOOLS WATER COMPANY	8104	1	60	1	1	4	T
1281	8303	1	HARTLEPOOLS WATER COMPANY	8204	1	60	1	1	4	T
1281	8403	1	HARTLEPOOLS WATER COMPANY	8304	1	60	1	1	4	T
1281	8503	1	HARTLEPOOLS WATER COMPANY	8404	1	60	1	1	4	T
1281	8603	1	HARTLEPOOLS WATER COMPANY	8504	1	60	1	1	4	T
1521	8112	2	WREXHAM AND EAST DENBIGHSHIRE WATER COMP	8101	1	60	1	1	3	T
1521	8212	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO	8201	1	60	1	1	3	T
1521	8312	1	WREXHAM AND EAST DENRIGHSHIRE WATER CO	8301	1	60	1	1	3	T
1521	8412	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO.	8401	1	60	1	1	3	T
1521	8512	1	WREXHAM AND EAST DENBIGHSHIRE WATER CO	8501	1	60	1	1	3	T
1581	8109	2	MID-SUSSEX WATER COMPANY	8010	1	60	1	1	2	T
1581	8109	4	MID-SUSSEX WATER COMPANY	8010	1	60	1	1	2	T
1581	8209	1	MID-SUSSEX WATER COMPANY	8110	1	60	1	1	2	T
1581	8403	1	MID-SUSSEX WATER COMPANY	8210	1	60	1	1	4	T
1581	8503	1	MID-SUSSEX WATER COMPANY	8404	1	60	1	1	4	T
1581	8503	4	MID-SUSSEX WATER COMPANY	8404	1	60	1	1	4	T
1581	8603	1	MID-SUSSEX WATER COMPANY	8504	1	60	1	1	4	T
1781	8203	2	LEE VALLEY WATER COMPANY	8104	1	60	1	1	4	T
1781	8303	1	LEE VALLEY WATER COMPANY	8204	1	60	1	1	4	T
1781	8403	1	LEE VALLEY WATER COMPANY	8304	1	60	1	1	4	T
1781	8503	1	LEE VALLEY WATER COMPANY	8404	1	60	1	1	4	T
1781	8603	1	LEE VALLEY WATER COMPANY	8504	1	60	1	1	4	T
1891	8112	2	MERSEY DOCKS & HARBOUR COMPANY	8101	1	70	1	1	3	T
1891	8212	1	MERSEY DOCKS & HARBOUR COMPANY	8201	1	70	1	1	3	T
1891	8312	1	MERSEY DOCKS AND HARBOUR COMPANY	8301	1	70	1	1	3	T
1891	8412	1	MERSEY DOCKS AND HARBOUR COMPANY	8401	1	70	1	1	3	T

From the ‘content validation’ section…

“Attempts to compare the dataset with summary figures from the published reports were not helpful. It is not clear how the data in the reports was derived from the original dataset, although the numbers of available accounts for each year are of a similar order of magnitude to those in this dataset. “



- **GFAC** - the grossing factor applied to a record - contains the value 1 or 300 in almost every instance, and this is generally related to the company size. However, record 10035 contains the value 100, and record 10261 contained 300.000002. The latter is shown as 300 in the converted data file, where it is represented as an integer.
- There are four occurrences of bad dates in the **DATE-PUBLISHED** field; one of 31 Feb, two of 31 Sep, and one of 0 Oct.
- Record 4689 contains an invalid **AREA-CODE** code of 'A'.
- Record 16306 contains an invalid **NATIONALITY** code of '45'.
- 277 occurrences of **BOX-NO** used codes whose meanings cannot be determined by NDAD.

The Company Accounts Data

- More than one version in the wild
- Ours – direct from government
  - Authentic; well-documented; inaccurate
- Others – altered by economists, social scientists
  - Poor provenance; less documentation; more accurate

Observations

- Message – quality is in the eye of the beholder
- Improving one aspect of quality can damage another
- Some markets provide many versions of the same data
  - Room for more of this approach with research data?

The conjugation of quality

- **I** want data that is accurate
- **You** want data that is up to date
- **She** wants data that is comprehensive
- **They** want data that is free

We probably cannot all be happy at the same time

Terminology - OAIS

- **Consumer** – anyone able to access data in the archive
- **Designated Community** – the set of Consumers that the archive aims to serve
- The Designated Community need not be a single community with a single set of interests

The problem

- We all want high-quality data
- Every data repository believes that its curation processes add quality
- But – when we talk about quality we are talking about different things
- Some aspects of quality conflict with others
- What does this mean for curation processes?
- How do we maximise re-use potential?

Engineer’s mantra

FAST, GOOD, CHEAP – pick any two!

Current curation practice

- Only one consumer group catered for per repository
- One workflow applies one set of quality controls and produces one dataset out for each dataset in
- Quality measures are often not explicit and rarely generic
- Disciplines differ greatly from the generic – contrast (e.g.) WDS with Zenodo

Clarification?

- Documented fully in work by Wang & Strong
  - Beyond Accuracy – what data quality means to data consumers, Journal of Management Information Systems 12(4); 1996
- Analysis goes further than 'research data'
- Some researchers interested in data from outside the academy
- Some data in universities has other, non-academic uses
- Data moves between government, academic, commercial and public sectors

Wang & Strong’s quality dimensions

Intrinsic	**Believability; Accuracy; Objectivity; Reputation**
Contextual	Value-added; Relevancy; Timeliness; Completeness; Appropriate amount
Representational	Interpretability; Ease of understanding; Representational consistency; Concise representation
Accessibility	Accessibility; Access security

Research repositories focus on some dimensions and neglect others

Some of these are to do with the systems rather than the data

Caveats

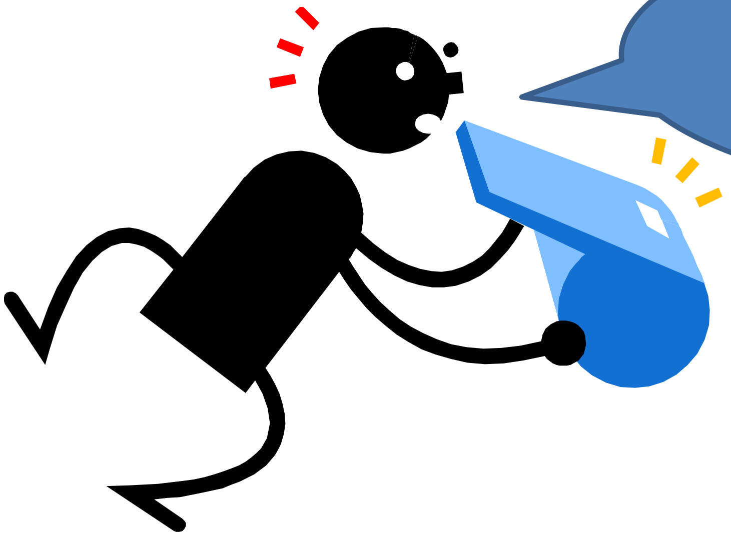
- Some important factors (e.g. cost) are hidden in these dimensions
- Some others are conflated (e.g. accuracy and precision)
- Not the full, or only, story – but any analysis is better than none

What can we gain?

- Greater mobility for data curation professionals – remove domain-specificity
- Increased number of generic data quality tools
- Training that emphasises transferable skills
- Ease of integration of data from disparate sources without error

Future curation, greater reuse

- Be explicit about quality metrics and curation processes in domain-independent ways
- Allow greater choice by Consumers
- Look harder at cost/benefit of quality processes – adjust where necessary
- Express quality in machine-readable interoperable ways



I need that data now!!! I don’t care how messy it is – I can fix it!

I’ve wasted too much of my life fixing other’s people’s bad data. I’m not interested until you’ve cleaned it up and documented it. Besides, I have other things to think about

County	Cereal Acres	LivestockAcres	Pop
Herts	23456	65345	.
Beds	12963	.	.
Bucks	42331	.	.
Essex	78113	.	.
Sussex	5428	.	.
Norfolk	99237	.	.

This value is highly accurate – to within .01%

This variable is highly precise – to 5SF

This table is incomplete – 10% of records are missing

Future curation & reuse

- Assertions automated, machine-readable
- Facilitates automated aggregation & analysis
- Big data emerges from the long tail
- Data released from sub-discipline silos
- Non-disciplinary repositories play a greater role
- Disciplinary archives can use expertise in wider domains
- Money is spent to best effect for research and reuse