Statistical Classifications: A FAIRy Tale

Faouzi Aloulou (EIA), Daniel Gillman (BLS),
Peter B. Meyer (BLS), and William Savino (Census)

Disclaimer: Views presented by the authors do not represent views of their agencies.

2024 FCSM Research and Policy Conference

October 22, 2024

Statistical classifications

Discrete categories represented by codes

- U.S. states and counties
- Industry
- Occupation
- Illness, injury, medical treatment
- Patent technology category

Data sets offered by the U.S. government use many of these.

- Our interest: Use & augmentation of classification variables in data sets
  - Mark data sets using a particular classification (data.gov)
  - Offer that info through a Web API
  - Machine-readable versions, ancillary tools
  - Share government practices
  - Create long time series, even if classifications have changed, via crosswalks or imputation

Enriched catalogs of classifications could help

FAIR principles

- 15 principles under these four guidelines/topics, supporting machine-actionable data and metadata
  - Findable
  - Accessible
  - Interoperable
  - Reusable

Motivation – Problem (1)

- In a data set, you encounter a column, with Industry Code, with values
    - 111140, 111150, 111160, 111211
- Are these from NAICS, SIC, or maybe even ISIC?
- Which version? Are multiple versions used?
- Are the codes provided valid? How to check?

Motivation – Problem (2)

- Previous example is with industry
- Same for occupation, product, disease, and others
- What & where are the referenced classifications?
- Which version is in use?
- Are there crosswalks to other versions and languages?
- Is there a file readable by a machine?
- Can we automate quality review?

Machine-actionable classification schemes

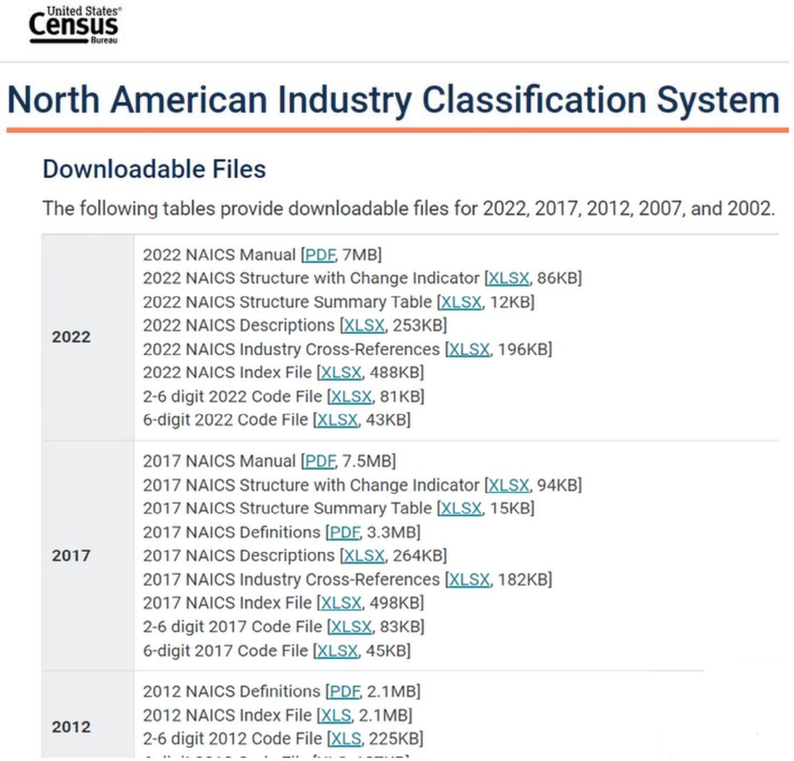
Classification management systems, or “classification servers” offer linked data on classifications, in different ways

- XKOS: an RDF vocabulary to publish classifications as Web Linked Data
- Colectica: uses DDI Lifecycle standard for classification management
- Aria: software for managing classifications used by Statistics Canada and Statistics New Zealand
- Schema.org offers formal metadata on classifications

These systems can manage crosswalks (correspondences) between classification systems.

Industry category systems

- NAICS is most common here.
  - Shared with Mexico and Canada
  - Census offers detailed crosswalks between NAICS versions
- Historically SIC, before 1997
  - Had different organizing principles
- NACE in EU
- ISIC internationally (UN)
- These are hierarchical
  - with hundreds of subcategories



United States Census Bureau

North American Industry Classification System

Downloadable Files

The following tables provide downloadable files for 2022, 2017, 2012, 2007, and 2002.

2022	2022 NAICS Manual [PDF, 7MB] 2022 NAICS Structure with Change Indicator [XLSX, 86KB] 2022 NAICS Structure Summary Table [XLSX, 12KB] 2022 NAICS Descriptions [XLSX, 253KB] 2022 NAICS Industry Cross-References [XLSX, 196KB] 2022 NAICS Index File [XLSX, 488KB] 2-6 digit 2022 Code File [XLSX, 81KB] 6-digit 2022 Code File [XLSX, 43KB]
2017	2017 NAICS Manual [PDF, 7.5MB] 2017 NAICS Structure with Change Indicator [XLSX, 94KB] 2017 NAICS Structure Summary Table [XLSX, 15KB] 2017 NAICS Definitions [PDF, 3.3MB] 2017 NAICS Descriptions [XLSX, 264KB] 2017 NAICS Industry Cross-References [XLSX, 182KB] 2017 NAICS Index File [XLSX, 498KB] 2-6 digit 2017 Code File [XLSX, 83KB] 6-digit 2017 Code File [XLSX, 45KB]
2012	2012 NAICS Definitions [PDF, 2.1MB] 2012 NAICS Index File [XLS, 2.1MB] 2-6 digit 2012 Code File [XLS, 225KB]

NAICS 2012

NAICS 2017

NAICS 2022

ISIC revs 0, 1, 2, 3, 3.1, 4

ISIC 5

NACE 1

NACE 1.1

NACE 1.2

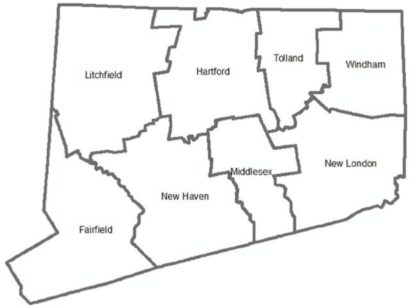
SIC 1982

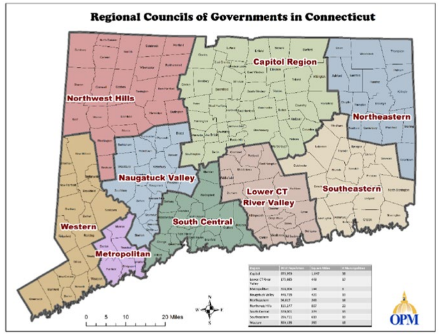
SIC 1987

Geospatial classification systems

- Geographical areas
  - Related to human-readable maps
  - Machine-readable shapefiles and polygons
  - Used for administration and survey methodology
- Challenge: change over time
  - Borders, jurisdictions, hierarchy

States
Counties
FIPS
Zip codes
hydrological areas
voting districts
ISO 3166
Census region
RUCA
PUMA
CBSA
MSA
SMSA
Sensor data
Census blocks





Connecticut’s 8 counties and 9 regions (Cassidy (2019) and Ratcliffe presentations)

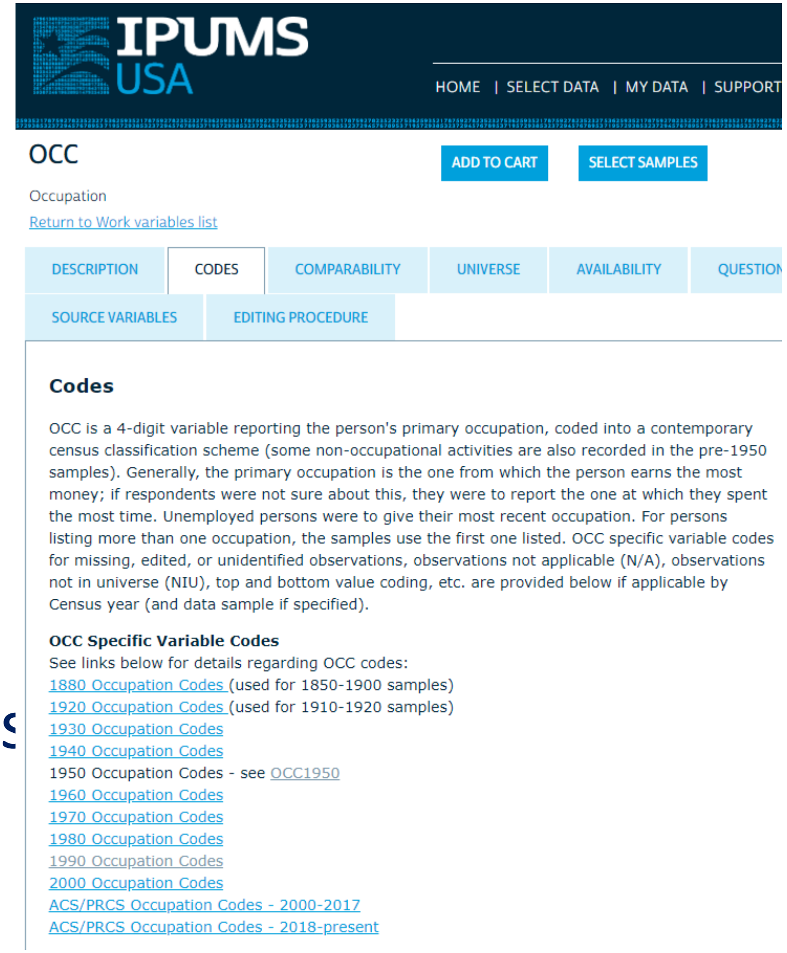
Occupations: Census and IPUMS

Univ. of Minnesota Population Center site offers augmented population Census data sets.

Includes original Census occupation in one variable (column) of Census or CPS or ACS data. These align to SOC over time.

Also offers standardized occupations from one year's classification for many other years (notably occ1950, occ1990, occ2010).

A data user chooses which ones to download.



OCC

Occupation

Return to Work variables list

Codes

OCC is a 4-digit variable reporting the person's primary occupation, coded into a contemporary census classification scheme (some non-occupational activities are also recorded in the pre-1950 samples). Generally, the primary occupation is the one from which the person earns the most money; if respondents were not sure about this, they were to report the one at which they spent the most time. Unemployed persons were to give their most recent occupation. For persons listing more than one occupation, the samples use the first one listed. OCC specific variable codes for missing, edited, or unidentified observations, observations not applicable (N/A), observations not in universe (NIU), top and bottom value coding, etc. are provided below if applicable by Census year (and data sample if specified).

OCC Specific Variable Codes
See links below for details regarding OCC codes:
1880 Occupation Codes (used for 1850-1900 samples)
1920 Occupation Codes (used for 1910-1920 samples)
1930 Occupation Codes
1940 Occupation Codes
1950 Occupation Codes - see OCC1950
1960 Occupation Codes
1970 Occupation Codes
1980 Occupation Codes
1990 Occupation Codes
2000 Occupation Codes
ACS/PRCS Occupation Codes - 2000-2017
ACS/PRCS Occupation Codes - 2018-present

Mapping between occupation classifications

A crosswalk is a mapping between discrete categories in one classification to categories in another. A crosswalk can usually be represented as a matrix.

Example: Occupations in Census 1990 and Occupations in Census 2000. The Census Bureau offers detailed crosswalks. They include percentages of each source category going into each destination category.

- Computer programmers to database administrators and Web designers

A crosswalk leaves out micro information; a more precise mapping can come from using many variables at once. Example on next slide.

Mapping between occupation classifications

A computer program could use more than two variables from the source data to find good matches for an observation in an external classification, using statistical modeling, or AI/ML

E.g. for a 1960 “lawyer or judge” one can use age, income, employer, location, etc. as predictors to impute “lawyer” or “judge” to each observation and thus split the category.

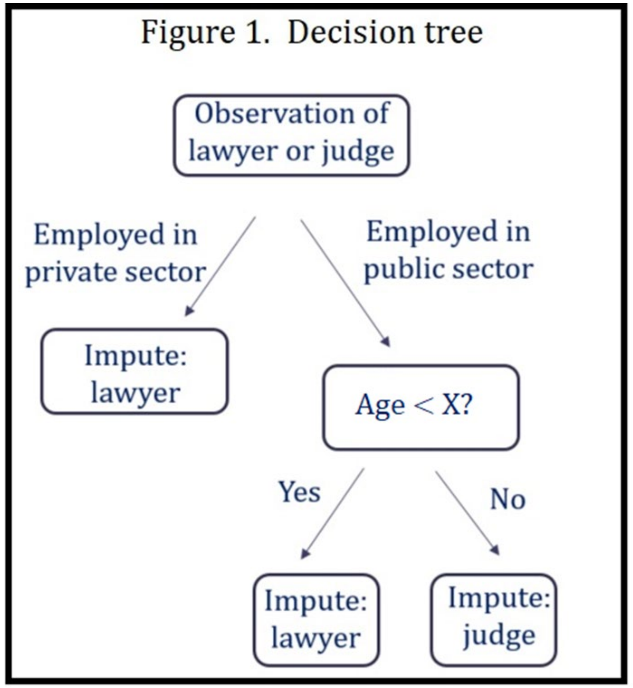


Figure 1. Decision tree

Observation of lawyer or judge
Employed in private sector
Employed in public sector
Impute: lawyer
Age < X?
Yes
No
Impute: lawyer
Impute: judge

Random forest-type decision tree for imputing flag lawyer or judge from three variables in random forest structure (Asher and Meyer 2021)

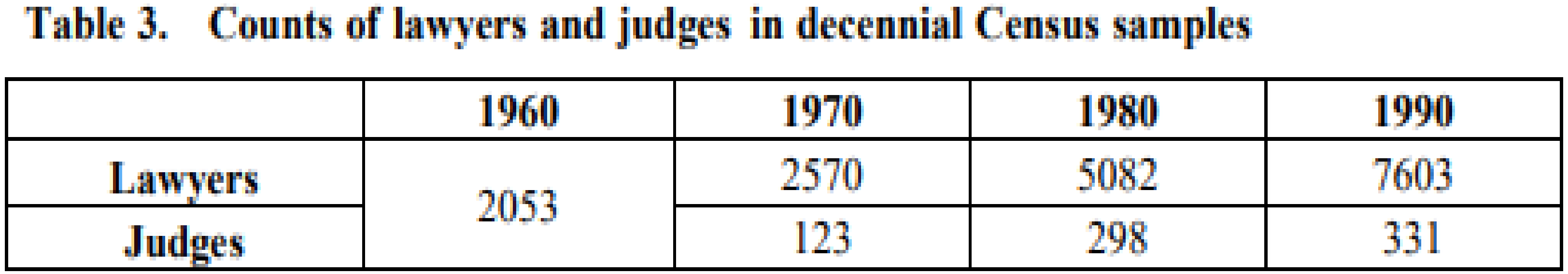


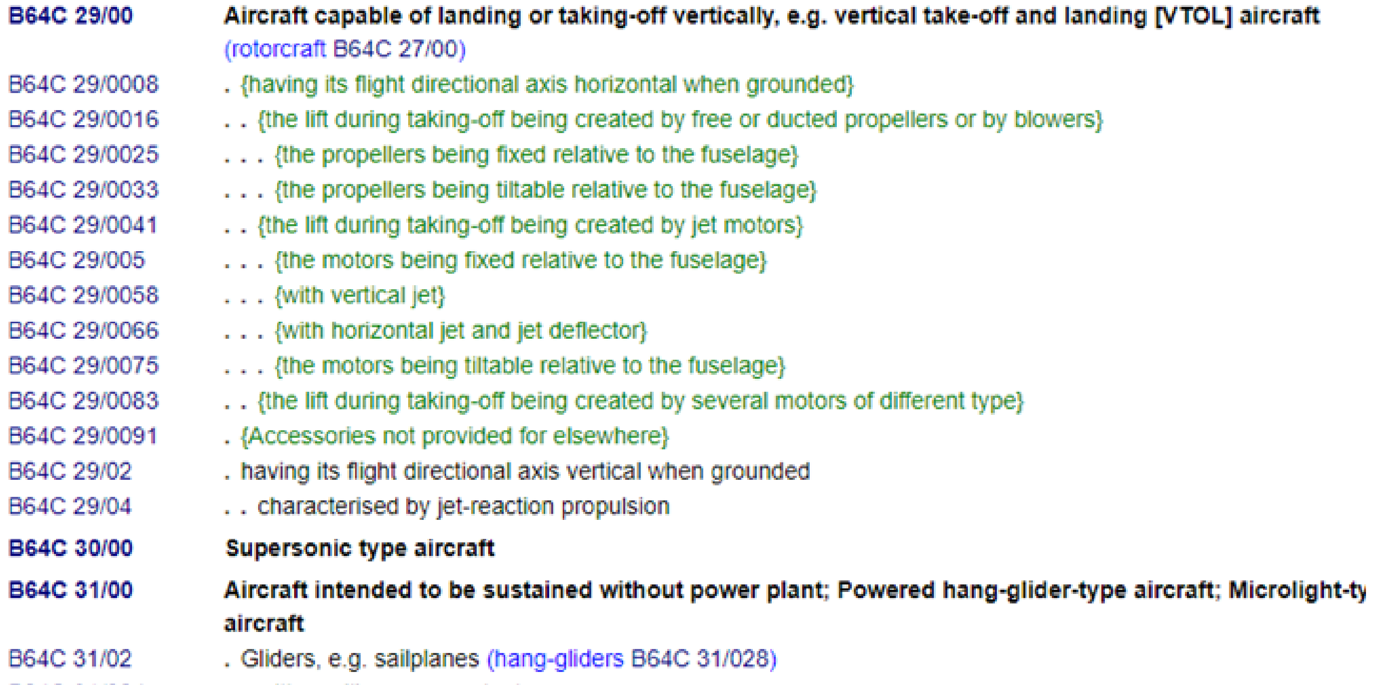
Table 3. Counts of lawyers and judges in decennial Census samples

	1960	1970	1980	1990
Lawyers	2053	2570	5082	7603
Judges		123	298	331

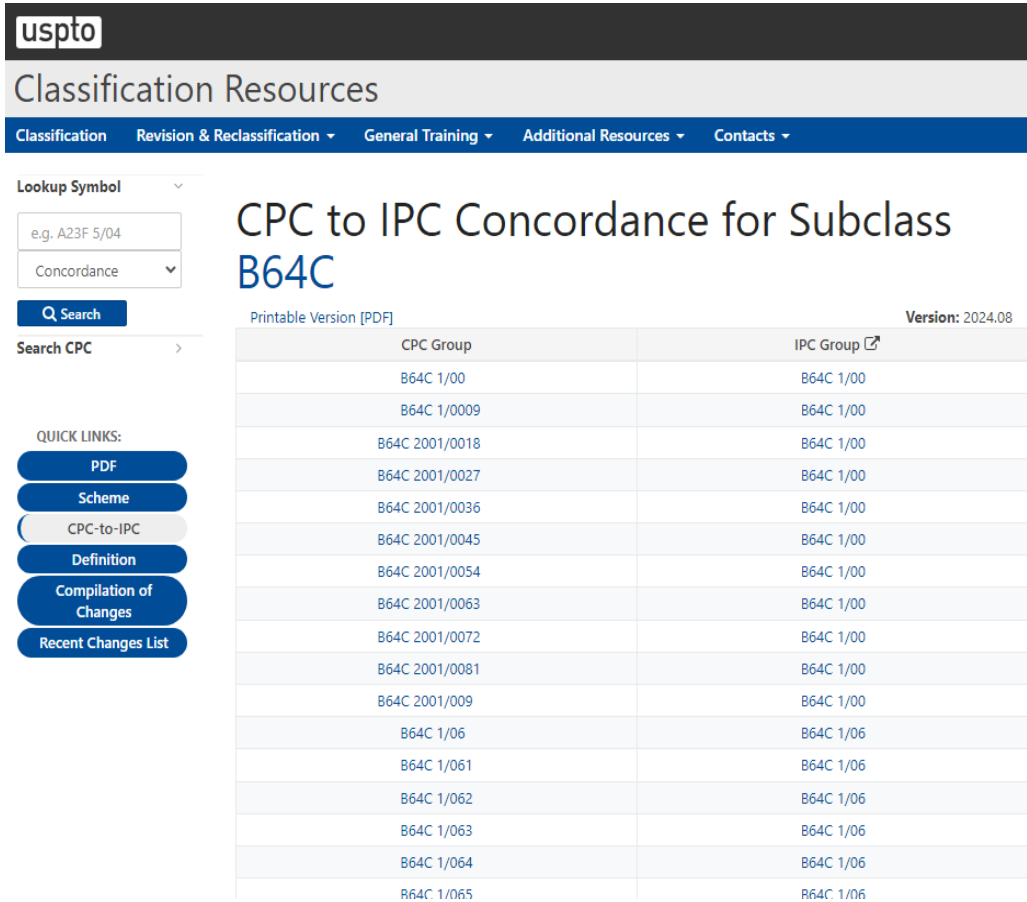
Patents classified by technology

There are many technology-classification systems for patents historically.

USPC, IPC, CPC, others in other countries



B64C 29/00	**Aircraft capable of landing or taking-off vertically, e.g. vertical take-off and landing [VTOL] aircraft** (rotorcraft B64C 27/00)
B64C 29/0008	. {having its flight directional axis horizontal when grounded}
B64C 29/0016	. . {the lift during taking-off being created by free or ducted propellers or by blowers}
B64C 29/0025	. . . {the propellers being fixed relative to the fuselage}
B64C 29/0033	. . . {the propellers being tiltable relative to the fuselage}
B64C 29/0041	. . {the lift during taking-off being created by jet motors}
B64C 29/005	. . . {the motors being fixed relative to the fuselage}
B64C 29/0058	. . . {with vertical jet}
B64C 29/0066	. . . {with horizontal jet and jet deflector}
B64C 29/0075	. . . {the motors being tiltable relative to the fuselage}
B64C 29/0083	. . {the lift during taking-off being created by several motors of different type}
B64C 29/0091	. {Accessories not provided for elsewhere}
B64C 29/02	. having its flight directional axis vertical when grounded
B64C 29/04	. . characterised by jet-reaction propulsion
B64C 30/00	**Supersonic type aircraft**
B64C 31/00	**Aircraft intended to be sustained without power plant; Powered hang-glider-type aircraft; Microlight-ty aircraft**
B64C 31/02	. Gliders, e.g. sailplanes (hang-gliders B64C 31/028)



CPC to IPC Concordance for Subclass B64C

Printable Version [PDF]

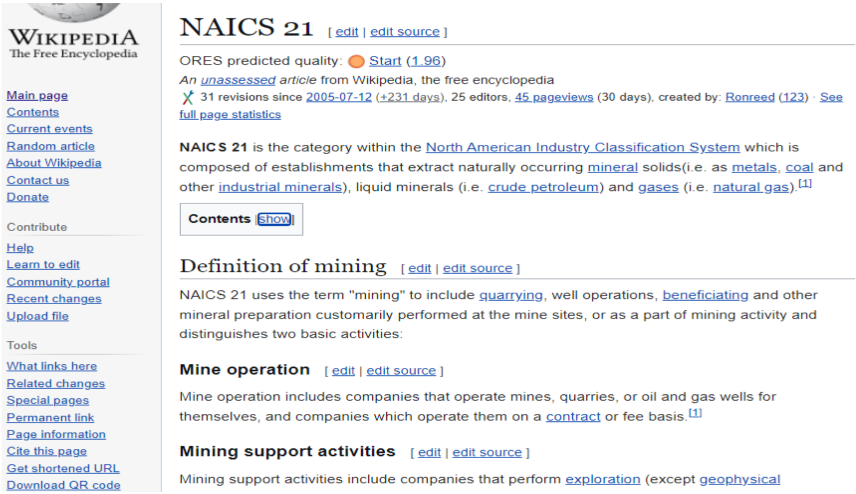
Version: 2024.08

CPC Group	IPC Group
B64C 1/00	B64C 1/00
B64C 1/0009	B64C 1/00
B64C 2001/0018	B64C 1/00
B64C 2001/0027	B64C 1/00
B64C 2001/0036	B64C 1/00
B64C 2001/0045	B64C 1/00
B64C 2001/0054	B64C 1/00
B64C 2001/0063	B64C 1/00
B64C 2001/0072	B64C 1/00
B64C 2001/0081	B64C 1/00
B64C 2001/009	B64C 1/00
B64C 1/06	B64C 1/06
B64C 1/061	B64C 1/06
B64C 1/062	B64C 1/06
B64C 1/063	B64C 1/06
B64C 1/064	B64C 1/06
B64C 1/065	B64C 1/06

Wikidata & Wikipedia can store classifications

- Wikidata entries are Linked Data and accessible by Web API; some have associated Wikipedia articles
  - Free and public
  - This approach allows citizen science, e.g. filling things in, crosswalks
  - A catalog could have detailed wiki pages with titles such as “Industry NAICS-2017-21”, on this or another wiki
  - A page about a classification/category could list or link to key terms, translations, crosswalks, predecessor categories, successors, parallels in other systems, e.g. translations across languages.

- Useful. It isn’t an official “controlled” vocabulary



NAICS 21 [edit | edit source]

ORES predicted quality: Start (1.96)

An unassessed article from Wikipedia, the free encyclopedia

31 revisions since 2005-07-12 (+231 days), 25 editors, 45 pageviews (30 days), created by: Ronreed (123) · See full page statistics

NAICS 21 is the category within the North American Industry Classification System which is composed of establishments that extract naturally occurring mineral solids(i.e. as metals, coal and other industrial minerals), liquid minerals (i.e. crude petroleum) and gases (i.e. natural gas).[1]

Contents [show]

Definition of mining [edit | edit source]

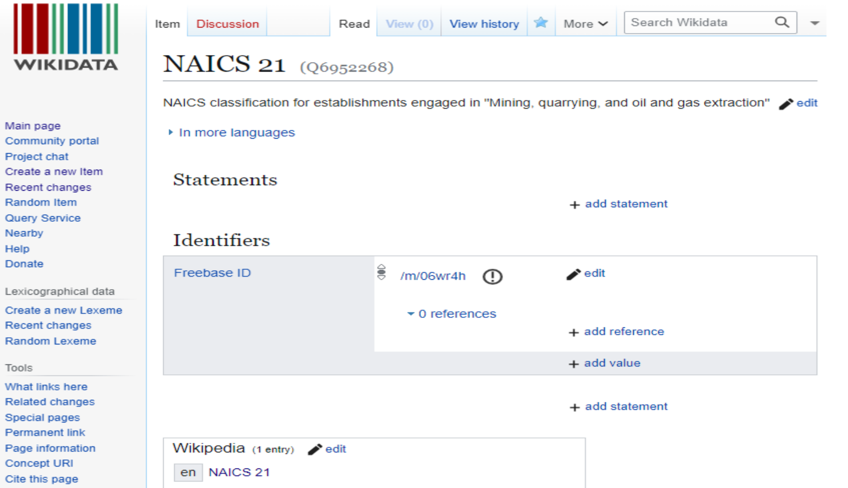
NAICS 21 uses the term "mining" to include quarrying, well operations, beneficiating and other mineral preparation customarily performed at the mine sites, or as a part of mining activity and distinguishes two basic activities:

Mine operation [edit | edit source]

Mine operation includes companies that operate mines, quarries, or oil and gas wells for themselves, and companies which operate them on a contract or fee basis.[1]

Mining support activities [edit | edit source]

Mining support activities include companies that perform exploration (except geophysical



NAICS 21 (Q6952268)

NAICS classification for establishments engaged in "Mining, quarrying, and oil and gas extraction" edit

In more languages

Statements

+ add statement

Identifiers

Freebase ID /m/06wr4h edit

0 references

+ add reference

+ add value

+ add statement

Wikipedia (1 entry) edit

en NAICS 21

Goals of this research

- Envision catalogs of statistical classifications
- Move toward FAIR goals regarding classification services
- Develop a broad understanding of this issue in the statistical community
- Describe solutions
- Build a prototype system (https://econterms.net/dg)

Conclusions

We don’t have a general machine-readable catalog meeting FAIR principles.

- Web sites provide a lot of information addressing statistical classifications
- Catalog services could help apply classifications to data
- Interpret classifications in data, find data using certain classification systems
- Services will help translate and map between classifications.
- Concordances may be increasingly machine readable – FAIR – for AI/ML systems

Contact

Peter B. Meyer

Research economist

Office of Productivity and Technology

U.S. Bureau of Labor Statistics

Meyer.peter@bls.gov

bls.gov/dpr/authors/meyer.htm