Metadata Application Profile

for

Geocoded Historical Industrial Sites Extracted from City Directories and Manufacturing Registries

Developed by "Developing a Spatial Approach for Toxic Transferal from Industrial and Vacant Land Uses to Green Infrastructure (GI)" (Brown, Texas A&M, and UCSD)[1]

Version Number: 1.1

Date Coverage: 2019-2020

DOI: https://doi.org/10.26300/9kj6-be70

[1] This project is funded by the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) under an administrative supplement P42 ES013660-14S1

Introduction

This document is a metadata application profile for geocoded historical industrial sites extracted from scanned text of city directories and manufacturing registries developed for the NIEHS-funded project "*Developing a Spatial Approach for Toxic Transferal from Industrial and Vacant Land Uses to Green Infrastructure (GI)*."

A metadata application profile is a set of recorded decisions about a shared data target for a given community[2]. The aim of this metadata profile is to recommend uniform content, use of authorities, and formats for the description of databases of geocoded historical industrial sites extracted from the scanned text of city directories and manufacturing registries using *georeg*[3] or *directoreadr*[4] software in order to both promote and increase the potential of their discovery and use, such as a map layer for visualization and analysis of historical industrial sites and their potential for identifying and tracking movements of soil and water contamination. In addition, the aim is to have data in open and common formats be deposited in a *FAIR-Enabling Repository* (as designated by the American Geophysical Union (AGU) and Coalition on Publishing Data in the Earth and Space Sciences (COPDESS), for the Environmental, Earth and Spaces Sciences communities[5] and/or EU's FairsFair project[6]. Currently libraries and archives use a variety of metadata standards to describe digital objects, including Metadata Object Description Schema (MODS), DublinCore, Data Documentation Initiative (DDI), and Schema.org. Geographic information is also currently described using more than one standard, such as the "ISO 191xx", e.g, ISO 19115 (2003) and ISO 19139-2:2012 and FGDC CSDGM. Furthermore, there are digital collection and discovery initiatives combining elements of existing standards to comprise a new metadata standard (e.g., OpenGeo and GeoBlacklight[7]). The deliverable of this part of the project is to ensure the metadata of files of databases of geocoded industrial sites extracted from scanned text of city directories and manufacturing registries are machine-actionable for use by researchers using data analysis and visualization platforms such as the *Data Discovery Studio*[8] developed at UCSD, which indexes data using ISO 19115 and Schema.org metadata and Esri.

---

[2] U.S. Library of Congress PCC Task Group on Metadata Application Profiles (2019) https://www.loc.gov/aba/pcc/taskgroup/Metadata-Application-Profiles.html

[3] georeg GitHub repository https://github.com/brown-ccv/georeg

[4] directoreadr GitHub repository https://github.com/brown-ccv/directoreadr

[5] COPDESS. Enabling FAIR Data Project (2019) http://www.copdess.org/enabling-fair-data-project/

[6] DataCite Repository Finder for Fair Enabling and EU FairsFair https://repositoryfinder.datacite.org/

[7] Barrista et al (2018) https://doi.org/10.1080/19386389.2018.1443414

[8] UCSD Data Discovery Studio. http://datadiscoverystudio.org/geoportal/

MAP Geocoded Historical Industry Data

Draft MAP below is based on Schema.org Types and Properties [9]

LABEL	DEFINITION	USAGE	PROPERTY
“name”	Dataset title	Include City Name and Year (ex. Industrial Site from Providence 1940 -1941)	Thing > CreativeWork > Dataset

LABEL	DEFINITION	USAGE	PROPERTY
“creator”	Dataset creator names	Lastname, Firstname M.I. @id “ORCID” “identifier” “https://orcid.org/0000-0002-7368-885X” (ORCID https://orcid.org/)	Thing > CreativeWork > Dataset

LABEL	DEFINITION	USAGE	PROPERTY
“affiliation”	Institution name	@id “ISNI” “identifier” “http://www.isni.org/isni/0000”(Institution Identifier (International Standard Name Identifier (ISNI)) http://www.isni.org/	Thing > Property > memberOf > affiliation (Note: sourceOrganization (Thing > CreativeWork > Dataset))

[9] Schema.org https://schema.org/docs/schemas.html

LABEL	DEFINITION	USAGE	PROPERTY
“spatialCoverage”	Location of sites in dataset	@id “ISO 3166-2” “State” “ISO 3166-1” “Country”	Thing > Property > contentLocation > spatialCoverage

LABEL	DEFINITION	USAGE	PROPERTY
“identifier”	Dataset’s Digital Object Identifier	Lastname, Firstname M.I. @id “doi” “identifier” “https://doi.org/x.xxxx/xxxxxx” (DOI System https://www.doi.org/)	Thing > Property > identifier

LABEL	DEFINITION	USAGE	PROPERTY
“type”	Type of digital object	@type “dataset”	Thing > Property > identifier

LABEL	DEFINITION	USAGE	PROPERTY
“inLanguage”	Language code	ISO 639-2b (text) IETF BCP 47 standard	Thing > CreativeWork

LABELS	DEFINITION	USAGE	PROPERTY
“license” “rights	permissions / rights status	(URL) Creative Commons “https://creativecommons.org/licenses/by-nc/4.0/” Rightstatements.org “http://rightsstatements.org/vocab/NoC-US/1.0/”	Thing > CreativeWork > Dataset

LABEL	DEFINITION	USAGE	PROPERTY
“funder”	Funder name and award identifier	Crossref controlled lost of funder names https://github.com/Crossref/open-funder-registry @id “ISNI” “identifier” “http://www.isni.org/isni/0000”(Institution Identifier (International Standard Name Identifier (ISNI)) http://www.isni.org/	Thing > CreativeWork > Dataset

LABEL	DEFINITION	USAGE	PROPERTY
“dateCreated”	Date the geocoded dataset was created (**NOT** the date of the registry/directory or its coverage)	w3cdtf ISO 8601 (YYYY-MM-DD)	Thing > Property > dateCreated

LABEL	DEFINITION	USAGE	PROPERTY
“temporalCoverage”	Time coverage/panel of the geocoded dataset ((Date Begin / Date End)	w3cdtf ISO 8601 (YYYY-MM-DD) Ex. 1940-1942 is expressed “1940/1942”	Thing > Property > dateCreated > https://schema.org/temporalCoverage

LABEL	DEFINITION	USAGE	PROPERTY
“latitude” “longitude”	Dataset Projection/Coordinate System	X, Y World Geodetic System (WGS84)	Thing > Intangible > StructuredValue > GeoCoordinates

LABEL	DEFINITION	USAGE	PROPERTY
“box”	Bounding box	X, Y World Geodetic System (WGS84) Ex. Rhode Island -71.8936578440155, 41.145497424786 1 -71.1199365422965, 42.018807606301 1	Thing > Intangible > StructuredValue > GeoShape > https://schema.org/GeoShape

Controlled Vocabularies and Ontologies

“Keywords” (Thing > Property > keywords)

Geographic: U.S. Geographic Names

- Geographic Names Information System (GNIS) (http://geonames.usgs.gov)

Geographic: International Geographic Names

- GEOnet Names Server (GNS) http://earth-info.nga.mil/gns/html/index.html

Topic: Dataset Chemical Names and Identifiers

- PubChem ID (CID) (ChemIDPlus and Hazardous Substances Data Bank now incorporated into PubChem)
- ChEBI ID
- Example: PFOS
    - Perfluorooctanesulfonic acid PubChem CID 74483 https://pubchem.ncbi.nlm.nih.gov/compound/74483
    - Perfluorooctane-1-sulfonic acid CHEBI:39421https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:39421
    - “Local” PFOS

Topic: Identifier for Dataset Specific Industry Names and Identifiers
URL

- NAICS (The North American Industry Classification System or NAICS) (2007,2012, 2017)
    - Gasoline stations (447110)
- SIC (Standard Industrial Classification (SIC)) (1987)
    - Gasoline service stations (5541) https://www.osha.gov/pls/imis/sicsearch.html

Topic: Dataset Subject Headings (Medical Subject Headings (MeSH))

- Environmental Health ID: D004782
- Cities ID: D002947
- Industry ID: D007221
- Manufacturing Industry ID: D066192
- Industrial Development ID: D000067896
- Chemical Industry ID: D002617
- Textile Industry ID: D013783
- Extraction and Processing Industry ID: D019652
- Metallurgy ID: D008669
- Oil and Gas Industry ID: D000066388
- Mining ID: D008906
- Manufactured Materials ID: D008420
- Jewelry ID: D055498
- Textiles ID: D013784
- Elastomers ID: D020724
- Power Plants ID: D011210
- Specialty Uses of Chemicals ID: D020313
- Environment and Public Health ID: D004778
- Hazardous Substances ID: D015386
- Toxic Actions ID: D004786
- Hazardous Waste ID: D006256
- Hazardous Waste Sites ID: D064371
- Industrial Waste ID: D007220
- Environmental Pollutants ID: D004785
- Costs and Cost Analysis ID: D003365
- Environmental Pollution ID: D004787
- City Planning ID: D002958
- Environmental Exposure ID: D004781
- Soil Pollutants ID: D012989
- Water Pollutants ID: D014873
- Air Pollutants ID: D000393
- Inorganic Chemicals ID: D007287
- Organic Chemicals ID: D009930
- Noxae ID: D009676
- Endocrine Disruptors ID: D052244
- Carcinogens ID: D002273
- Environmental Restoration and Remediation ID: D052918
- Urbanization ID D014507
- Environment ID: D004777
- Ecological and Environmental Phenomena ID: D055669
- Ecological Parameter Monitoring ID: D064308
- Ecosystem ID: D017753
- Urban Health ID: D014504
- Urban Population ID: D014505
- Social Conditions ID: D012924
- Socioeconomic Factors ID: D012959
- Water Supply ID: D014881
- Groundwater ID: D060587
- Floods ID: D055868
- Environment Design ID: D004779
- Demography ID: D003710
- Conservation of Natural Resources ID: D00324

Related Items

Related Publication Citations and Related Items (E.g., Physical/Digital Registries and Directories)

Directory/Registry Title
Directory/Registry Location
Directory/Registry Copyright Date

- w3cdtf ISO 8601 (YYYY-MM-DD)

Directory/Registry Publisher
Directory/Registry Publisher Location
Directory/Registry Volume
Directory/Registry Identifiers

- (ISBN, DOI, ARK, OCLC)

Directory/Registry Language

- Iso639-2b (text)

Directory/Registry URL

Title: Polk's Providence (Providence County R.I.) and Rhode Island State Business Directory
Volume: 1940
Copyright date: 1940
Publisher Name: R.L. Polk and Company
Publisher Location: 509 Westminster Street, Providence, Rhode Island | 179 Lincoln Street Boston Massachusetts
Identifier-ark ark:/13960/t28980d29
OCLC Number: 1050798059
URI: https://www.worldcat.org/title/polks-providence-providence-county-ri-city-directory/oclc/1050798059
URL: https://archive.org/details/polksprovidencep1940unse/page/n9/mode/2up

Format Curation

The native tabular data format should be retained and a copy exported into an open CSV format for digital preservation best practices. For making data actionable the tabular data should be reformatted into tables using an open GIS mapping software such as QGIS or geodatabase feature class using Esri GIS mapping software.

Recommended metadata best practice should include complete information on:

- Bounding box
- Geographic location (place keywords)
- Spatial representation type (vector, raster, etc.)
- Projection/coordinate system

For information on this curatorial best practice for geodatabases[10] and geoJSON see Data Curation Network (DCN) primers.

[10] Data Curation Network (DCN). GeoDatabase Data Curation Primer (2020). https://datacurationnetwork.org/resources/data-curation-primers/