Database citeseerx: Stores all information about papers metadata.

papers: Stores metadata and access information for each paper.

Field	Description
id	Unique id identifying each paper. System assigned using the DOI server.
version	last valid version of paper metadata
cluster	Identifier of the collection of similar papers/citations. Default value 0, after the inference process is run.
public	Indicates if the paper is available or not
ncites	Number of citations found within the corpus for the given paper. (It's calculated by the inference process)
versionName	Name of the last valid version of paper metadata
crawlDate	When the paper was obtained
repositoryID	Repository identifier where all the files associated to the paper are located. The Repository identifier maps to a physical location in a file system
ConversionTrace	All the tools (in order) used to get the text version of the paper.
selfCites	Number of self citations (This data is calculated by the inference process)
versionTime	When the last version of the paper metadata was created/updated

papers_versionShadow: Stores information indicating where the last version of the paper metadata came from (Header Parser, Inference, User Corrections). Each field stores the name of the source for the same field in the papers table.

authors: Information about each one of the authors for a given paper

Field	Description
id	Unique id identifying for each author. System assigned auto increment.
cluster	Groups name variations for the same author. The identifier is assigned by the name disambiguation process
name	Authors name as extracted/corrected from the paper
afill	Affiliation as extracted/corrected from the paper
address	Author's Address as extracted/corrected from the paper
email	Author's email address as extracted/corrected from the paper
ord	Location of the author in the list of author for the given paper

authors_versionShadow: Stores information indicating where the last version of the author metadata came from (Header Parser, Inference, User Corrections). Each field stores the name of the source for the same field in the authors table.

acknowledgments: Acknowledgments found on a paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
cluster	Identifier for similar/same acknowledgment
name	
ackType	

acknowledgments_versionShadow: Stores information indicating where the last version of the acknowledgment metadata came from (Header Parser, Inference, User Corrections). Each field stores the name of the source for the same field in the acknowledgment table.

acknowledgmentContexts: Contexts in which an acknowledgments was found inside a paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
context	Raw text where the acknowledgment was found

citations: Citations found on a paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
cluster	Unique identifier that match the same citation in different papers to a canonical one; including the paper if present within the corpus. The id is assigned by the inference process
raw	Citation text as found on the paper
self	Indicates if this citation is a self citation

citationContexts: Contexts in which a citation was found inside a paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
context	Raw text where the citation is mention on the paper

keywords: Keywords found in the paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
keyword	The keyword

keywords_versionShadow: Stores information indicating where the last version of the keyword metadata came from (Header Parser, Inference, User Corrections). Each field stores the name of the source for the same field in the keywords table.

checksum: Checksum data for each file type associated to a paper. A paper can have different file type representations like PDF, PS, and so on. The same paper could also have different versions in a same file type; let's say the PDF pre-print and camera ready versions of the paper.

Field	Description
sha1	Digest of the paper calculated using the SHA1 algorithm
fileType	Type of the file. For intance, PDF, PS.

paperVersions: Information about each one of the metadata corrections for a given paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
name	Assigned name for that version
version	Number of this version (1st, 2nd, and so on)
repositoryID	Repository identifier which maps to the physical location where the version information is stored
path	Relative path from the ROOT of the repository to the location of the version information
deprecated	Flag to indicate if the version has been deprecated or not
spam	Flag indicating if the version is spam or not
time	When the version was created

userCorrections: Information about a metadata correction made by a user

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
version	The version number

urls: URL where a version of the paper was found. A paper can be found at different URLs. For example, a URL pointing to a PDF version of the paper, another one pointing to the PS version within the same web server, and / or different URLs pointing to the same PDF version but in different web servers one for each one of the paper's authors.

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.

url	URL to the paper

hubUrls: URLs pointing to web pages which contains links to PDF, PS files

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
url	URL to a resource which contains links to PDF/PS documents
lastCrawl	Last time when this URL was crawled
repositoryID	

hubMap: Associates urls (pointing to a paper) with its page container if any

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.

legacyIDMap: Mappings beetwen citeseer and citeseerx papers (CiteSeerX table only)

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.

textSources: Used to set up a text source (right now only the banner).

Field	Description
name	Text source name.
content	Content of the text source

tags: Tags associated to a paper

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
tag	The tag
count	Number of times this tag has been used

link_types: Identifies external sources that can be linked for each paper.

Field	Description

label	External source identifier.
baseURL	Base URL to the external source

elinks: External links for papers based on information from the paper, link_types, and an external metadata database.

Field	Description
label	Link type idenfier
url	Complete URL pointing to information about the paper on an external source. The URL is built base on the baseURL from link_types and a process that match papers within the corpus with the external sources

pdfTables: Metadata and data from each one of the tables found in a paper

Field	Description
id	Unique id identifying for each acknowledgment. The id is based on the paperid (Note by Juan: I think this should be an auto increment)
caption	Table caption text
content	Table content in HTML format. (Note by Juan: This should be stored in XML and probably in the file system or both the database and file system)
footNote	Table foot note
refText	Text snippets where the table is mentioned. (Note by Juan: This should be treated the same way citation and acknowledgment context are handled)
pageNum	Page where the table was found within the paper

cannames: Cannonical/disambiguated author names information. (could we name this as canonical_authors?)

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
sid	
canname	Cannonical Name for an author
ndocs	Number of papers within the corpus for that author
ncites	Number of citations for the author within the corpus
url	Home page for the author
affil	Author's current affiliation
address	Author's current address
email	Author's current email address
hindex	Author's hindex

citecharts: Need to figure it out what this table is storing

Database csx_citegraph: Stores all citation graph information. Data in this database is maintained by the inference process.

clusters: Stores the canonical reference for paper and/or citations.

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
size	Cluster size
incollection	Flag to indicate if the reference is a paper within the corpus or just a citation
observations	
selfCites	Number of seltcitations
updated	Last time the cluster was updated. For instance, adding a new observation

keymap: Stores several keys to allow citation/paper matching. Keys are generated using normalization over the title and authors names.

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
ckey	The generated key

citations: a Stores all the citations for a given paper which have been assigned to a cluster. (If a citation appears here it must also appear in citeseerx.citations)

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
cluster	The cluster this citation belongs to

papers: Stores all the papers which have been assigned to a cluster. (If a paper appears here it must also appear in citeseerx.papers)

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
cluster	The cluster this paper belongs to

citegraph: Associates papers with: citing and cited clusters

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.

citing	The citing cluster
cited	The cited cluster
firstContext	The first context where the citation is found
selfCite	Flag which indicates if this one is a self citation

infupdates: ?

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
lastupdate	

deletions: Stores deleted clusters so they are deleted from the index?

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
deleted	

IndexTime: Last time the indexer was run

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
param	Process name
lastupdate	Last time the process was run

Database csxdoi: Handles Digital Object Identifiers assignment

doi_granted: Stores data about granted Digital Object Identifiers

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
doi_type	Digital Object Identifier type (Paper, Hub URL, File, and so on)
bin	
rec	
assigned	Time when this DOI was assigned

configuration: DOI Server configuration

Field	Description
deployment	

site_id	Site Identifier
deployment_id	Deployment identifier. A site can have many deployments

Database myciteseerx: Personal portal database for main site

users: Basic user information

Field	Description
userid	User identifier
password	User password. Encoded
firstName	User first Name
middleName	User middle name
lastName	User last name
affil1	Users primary affiliation (Organization)
affil2	Users secondary affiliation (department within organization)
enabled	Flag which indicates in the user is enabled or not
country	User's country
province	User's province
webPage	User's web page
intenalid	System assigned identifier
updated	Last time information about this user changed

authorities: Rights assigned to users (Spring Security Integration)

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
authority	Name of the authority given to the user

SubmissionJobs: Stores information about URL's send by users containing papers

Field	Description
JID	Job identifier
UID	User which send the job
URL	The URL to be crawled
depth	How many hops the crawler should make from the URL. Default 1
time	Time when this job was posted
status	The Job status

statusTime	Last time the status for this job was updated

submissionComponents: Stores information about each one of the component of a submission job if any. If the submission Job is a hub URL, a HTML page which contains links to PDF/PS files, a record for each link is stored in this table.

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
JID	Job identifier
URL	URL to the component
status	Status of this component
time	
DID	

activation: Activation codes used to create user accounts

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
code	Activation code
created	When the activation code was created

Invitations: Invitations tickets

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
ticket	

configuration: configuration parameters

Field	Description
param	Configuration paramater
value	Value of the parameter

collections: user created collections. Initially intended to store papers

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
name	Collection name
description	Collection description

collection_notes: Notes the user have associated to a particular collection

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
CID	Collection Identifier
UID	User id
note	The note

papers_in_collection: The papers the user has put in a particular collection. All the papers ids come from citeseerx.papers

Field	Description
CID	Collection Identifier
PID	Paper Id
UID	User id

paper_notes: Notes a user has done to a particular paper in a specific collection

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
CID	Collection id
PID	Paper id
UID	User id
note	The note

monitors: Stores information about papers the user is monitoring. Mainly, changes in metadata

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
userid	User monitoring a paper
paperid	The paper being monitored

tags: User created tags for a paper

Field	Description

id	Unique id identifying for each acknowledgment. System assigned auto increment.
userid	The user that created the tag
paperid	The tagged paper
tag	Tag text
added	When this tag was added

groups: Collection of users

Field	Description
id	Unique id identifying for each acknowledgment. System assigned auto increment.
name	Group name
description	Group description
owner	User who created the group
authority	Authorithy identifier assigned to the group (to handle permissions using the group)

group_members: Stores the users that part of a particular group

Field	Description
groupid	Group
userid	User
validating	Indicates if the user needs to be validated to have access to the this group

IndexTime: Last time a indexable object in this database was indexed

Field	Description
param	Indicates the entity
lastTime	Indicates the last time this entity was indexed

ACL_* tables are used to stored authorization rights to authorities (users, groups) over domain objects. Authorization is handled by Spring Security.

Database: csx_external_metadata

This database stores external metadata that can be used by several process within the system. Right now, external metadata from DBLP and CiteULike is stored.

CiteULike metadata and part of the DBLP metadata is used to create external links for each paper on the system corpus.

Metadata from DBLP can also be used to correct records, or disambiguate authors names

This part of the system is semi-generic. However the external metadata is dependent of the system being build. CiteULike metadata can be useful for a Chemistry system but DBLP metadata probably not.