Chapter 13

NDEx: A Community Resource for Sharing and Publishing of Biological Networks

Rudolf T. Pillich, Jing Chen, Vladimir Rynkov, David Welker, and Dexter Pratt

Abstract

Networks are a powerful and flexible paradigm that facilitate communication and computation about interactions of any type, whether social, economic, or biological. NDEx, the Network Data Exchange, is an online commons to enable new modes of collaboration and publication using biological networks. NDEx creates an access point and interface to a broad range of networks, whether they express molecular interactions, curated relationships from literature, or the outputs of systematic analysis of big data. Research organizations can use NDEx as a distribution channel for networks they generate or curate. Developers of bioinformatic applications can store and query NDEx networks via a common programmatic interface. NDEx can also facilitate the integration of networks as data in electronic publications, thus making a step toward an ecosystem in which networks bearing data, hypotheses, and findings flow seamlessly between scientists.

Key words Biological networks, REST API, Java, Open source, BioPAX3, Big data, Pathway analysis, OpenBEL, Python, Cytoscape, Provenance, Cyberinfrastructure

1 Introduction

Networks are a precise and computable form in which biologists can express many kinds of information, including models of biological mechanisms, experimental facts, and relationships derived by systematic data analysis. When pathway diagrams evolved into repositories of small pathway networks [1, 2, 3], they became searchable resources and the basis for data interpretation and collaborative pathway editing applications [4]. In recent years, the development and progress in omics technologies have contributed to boosting the construction of networks inferred by the systematic processing of big data, providing an important avenue for interpretation and a counterpoint to literature curation [5, 6, 7]. In this view, it is necessary to create an infrastructure where the rapidly expanding corpus of biological network models created by

Cathy H. Wu et al. (eds.), *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, Methods in Molecular Biology, vol. 1558, DOI 10.1007/978-1-4939-6783-4_13, © Springer Science+Business Media LLC 2017

researchers can be stored, shared, discussed, reviewed, and used. This chapter describes NDEx, the Network Data Exchange, an online commons where scientists can store, share and publicly distribute biological networks as dynamic actionable data, and develop applications using them [8]. One of the goals of the NDEx Project is to create a home for network models that are currently available only as figures, tables, or supplementary information, such as networks produced via systematic mining and integration of large-scale molecular data. In doing this, the NDEx project does not compete with existing pathway and interaction databases, such as KEGG or Reactome; instead, NDEx provides a novel, common distribution channel for these efforts, preserving their identity and attribution rather than subsuming them.

By providing a flexible computable medium for biological knowledge, networks are also becoming a critical element for new models of scientific publication, in which data and its derivatives are as important as text [9]. The NDEx platform is intended to enable experimentation with novel forms of scientific review and discourse. NDEx networks are assigned stable, globally unique URIs and can be referenced by publications, by other networks, and by analytic applications.

NDEx aims to become the main hub for the development of new, lightweight applications and scripts capable of accessing and manipulating networks via the NDEx API, making it easy for scientists to develop novel network-based analyses. An example is CyNDEx, the NDEx Cytoscape App: CyNDEx enables users to access an NDEx server directly from Cytoscape and engage its full range of tools to analyze and transform any networks stored on NDEx. The NDEx project fosters the creation of new utilities and analytic tools that use NDEx via code examples, client libraries in multiple languages, developer documentation, and a strong social media outreach campaign to drive community awareness and engagement.

2 Materials

2.1 NDEx Sources

All NDEx sources are stored on GitHub in publicly accessible repositories under the **ndexbio** organization: https://github.com/ndexbio.

The following repositories support released NDEx software. For all other repositories under ndexbio, we advise you to consult with the NDEx team on their status before using. Some are in active development, while others may be obsolete or highly experimental.

2.2 NDEx Server

https://github.com/ndexbio/ndex-rest

https://github.com/ndexbio/ndex-object-model

https://github.com/ndexbio/ndex-common

2.3 NDEx Web App

https://github.com/ndexbio/ndex-webapp

2.4 NDEx Java Client

The NDEx Java Client is not an application; it is a library available for use by developers to create Java applications that access NDEx.

https://github.com/ndexbio/ndex-java-client

https://github.com/ndexbio/ndex-object-model

2.5 NDEx Python Client

The NDEx Python Client is a library facilitating access to networks on NDEx servers from Python. It provides convenient methods to find, query, and save networks while managing authentication.

https://github.com/ndexbio/ndex-python-client

2.6 CXIO

Java library is used for CX serialization and de-serialization by the NDEx server and the NDEx Java Client library.

https://github.com/cytoscape/cxio

2.7 CyNDEx Cytoscape App

https://github.com/ndexbio/ndex-cytoscape-app

https://github.com/ndexbio/ndex-java-client

https://github.com/ndexbio/ndex-object-model

2.8 NDEx Sync Copier Utility

https://github.com/ndexbio/ndex-sync

https://github.com/ndexbio/ndex-java-client

https://github.com/ndexbio/ndex-object-model

3 What Is NDEx

3.1 Comparison of NDEx to Other Network Resources

NDEx is an online resource to enable collaboration and publication using biological networks. It is a "commons," a scientist-driven data exchange, where both individuals and organizations can share networks of any type, from pathway models and interaction maps in standard formats to a novel data-driven knowledge. This user-centric focus differentiates NDEx from the array of biological network resources currently available to biologists. In many cases NDEx is complementary to the missions of existing network resources, potentially playing roles as a novel distribution channel, a user content management component, or a source for staging of prepublication content. Most of these resources can be categorized either as **repositories** of network-structured information, as **analysis applications** that operate on input data (such as gene lists) via techniques that use one or more reference networks, or as **both**. "Repositories" in this context means resources where the network content is *managed by the organization maintaining the resource* and is therefore different from the structure of NDEx in which the *users manage the network content*. Some well-known examples of repositories include KEGG (http://www.genome.jp/kegg/pathway.html), Pathway Commons (http://www.pathwaycommons.org/about/), IntAct (http://www.ebi.ac.uk/intact/), and BioCyc (http://biocyc.org/). Many repositories also differ from NDEx because they use specific network formats and models of biology, in contrast to the NDEx strategy of supporting many formats in a common framework; one such example is the NCI/Nature—curated Pathway Interaction Database (http://pid.nci.nih.gov/) that only uses the BioPAX3 format.

Analysis applications using network resources include sites such as GeneMANIA (http://www.genemania.org/) and NCI DAVID (https://david.ncifcrf.gov/). Although NDEx provides some search and query operations that could be construed as "analysis," its mission is not to perform biological analyses but instead to be a service that facilitates the creation of applications, both as a source of reference networks and as a place for users to store network-structured analysis results. A recent example of a network-oriented analysis application is the Network Portal (http://networks.systemsbiology.net/) by the Institute for Systems Biology, which "provides analysis and visualization tools for selected gene regulatory networks to aid researchers in biological discovery and hypothesis development." Its design includes several features to promote data sharing and integration with other applications, but its primary focus is *analysis* using networks of transcriptional regulation.

WikiPathways [10] (http://www.wikipathways.org/index.php/WikiPathways) is a pioneering collaborative platform for the curation of biological pathways, a resource that shares the NDEx goal of facilitating scientific discourse by providing a platform for user-driven content. It differs, however, in that: (1) it is focused on pathway diagrams that are small and curated and in which the content may not be fully represented as a network, and (2) it employs the "Wiki" model of collaboration on a public document, different from the "Google Docs" approach of NDEx in which users manage the access to their networks. The role of NDEx in the context of collaborative environments such as WikiPathways could be as a "back end" resource to store and share the content created by the collaborators.

BioModels at EBI (https://www.ebi.ac.uk/biomodels-main/) is an example of a database of biological information that could be considered a network resource, but which is different from NDEx not only because the content is managed but also because it is specialized to a particular kind of biological data structure. BioModels is a "repository of computational models of biological processes," serving as resource for the computational modeling community. Although there are forms of these computational models that can be expressed as networks (and which

NDEx may support at some point), BioModels presents these models in a comprehensive manner tailored to the needs of its user community of bioinformatic experts.

To conclude, NDEx provides a novel distribution strategy for organizations that maintain repositories, a new channel for their content to reach users and applications.

3.2 Available Online Resources

The following sections present a comprehensive but not exhaustive list of network repositories and selected examples of network-oriented analysis applications. The repositories include both those that are based on curated mechanistic information ("pathways") and those that are focused on interaction data. PathGuide (http://www.pathguide.org/) was an invaluable resource for the preparation of this section.

3.2.1 Aggregators of Network Resources

Pathway Commons–http://www.pathwaycommons.org/about/

- "Pathway Commons is a network biology resource and acts as a convenient point of access to biological pathway information collected from public pathway databases, which you can search, visualize and download."
- Aggregator of network repository data from many sources.
- Normalizes resources to BioPAX3.
- Distributes in SIF and BioPAX3 formats.

iRefIndex–http://irefindex.org/wiki/index.php?title=iRefIndex

- "Provides an index of protein interactions available in a number of primary interaction databases including BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MINT, MPact, MPIDB, MPPI and OPHID."

3.2.2 Protein–Protein and Other Molecular Interaction Networks

Quoted sentences below and in Subheadings 3.2.4 and 3.2.5 are taken from the resources' websites.

BioGRID—http://thebiogrid.org/

- "BioGRID is an interaction repository with data compiled through comprehensive curation efforts."

CCSB Interactome—http://interactome.dfci.harvard.edu/

- A repository of experimentally derived protein interactions.

DIP—http://dip.doe-mbi.ucla.edu/dip/Main.cgi

- "The Database of Interacting Proteins (DIP™) catalogs experimentally determined interactions between proteins."

IntAct—http://www.ebi.ac.uk/intact/

- A central, standards-compliant repository of molecular interactions, including protein–protein, protein–small molecule, and protein–nucleic acid interactions.

- IntAct provides both an open-source database system and analysis tools for molecular interaction data and acts as common curation platform for 11 molecular interaction databases.

NetPro—http://www.molecularconnections.com/products/#netpro "NetPro™ is a comprehensive database of Protein–Protein and Protein–Small molecules interaction, consisting of more than 320,000 interactions captured from more than 1500 abstracts, approximately 1600 published journals and more than 60,000 references."

STRING—http://string-db.org/

- "STRING is a database of known and predicted protein interactions.

  The interactions include direct (physical) and indirect (functional) associations."

MINT—http://mint.bio.uniroma2.it/mint/Welcome.do

- "MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators" and is now integrated with IntAct.

RBPDB—http://rbpdb.ccbr.utoronto.ca/

- Repository of RNA–protein interactions.

BioLiP—http://zhanglab.ccmb.med.umich.edu/BioLiP/

- "BioLiP is a semi-manually curated database for high-quality, biologically relevant ligand-protein binding interactions."

BindingDB—http://www.bindingdb.org/bind/index.jsp

- "BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules."

TRANSFAC—http://www.gene-regulation.com/index2

- Commercial repository of gene regulation interactions, subset available for academic use.

iMEX—http://www.imexconsortium.org/

- "A non-redundant set of protein–protein interaction data from a broad taxonomic range of organisms." iMEX also provides access to a selection of data curated by other resources such as IntAct, MINT, and DIP.

TAP Project—http://tap.med.utoronto.ca/exttap/

- "The Yeast TAP Project is aimed at elucidating the entire network of protein-protein interactions in a model eukaryotic organism, namely, the yeast *Saccharomyces cerevisiae*."
- Repository derived from experimental data using tandem affinity purification (TAP).

3.2.3 Pathway Network Resources

NetPath—http://www.netpath.org/

- "'NetPath' is a manually curated resource of signal transduction pathways in humans."

NCI–PID—http://pid.nci.nih.gov/

- The NCI/Nature—curated Pathway Interaction Database is a collection of "Biomolecular interactions and cellular processes assembled into authoritative human signaling pathways."
- Officially retired as of December 31, 2015, the NCI–PID relies on NDEx as its primary distribution channel.

Reactome—http://www.reactome.org/

- "Reactome is a free, open-source, curated and peer reviewed pathway database whose goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology, and education."

SignaLink Database—http://signalink.org/

- "An integrated resource to analyze signaling pathway cross-talks, transcription factors, miRNAs and regulatory enzymes."

WikiPathways—http://www.wikipathways.org/index.php/WikiPathways

- "WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community."
- Pathway diagrams are sometimes only partially computable, incorporating graphic elements with meaning apparent to the biologist but difficult for algorithms to interpret.

BioCyc—http://biocyc.org/

- "BioCyc is a collection of 5711 Pathway/Genome Databases (PGDBs), plus software tools for understanding their data." It includes EcoCyc and MetaCyc and is focused on metabolism.

KEGG—http://www.genome.jp/kegg/pathway.html

- Repository of manually curated pathway and interaction networks and diagrams.

MANET—http://manet.illinois.edu/aboutManet.php

- "The Molecular Ancestry Network (MANET) database project traces evolution of protein architecture onto biomolecular networks."

SMPDB—http://smpdb.ca/

- "An interactive, visual database containing more than 618 small molecule pathways found in humans."
- Includes extensive, carefully formatted diagrams and exports in BioPAX3 and SBGN.

Atlas of Cancer Signaling Networks—https://acsn.curie.fr

- "ACSN is a pathway database and a web-based environment that contains a collection of interconnected cancer-related signaling network maps."
- Uses SBGN created with cell designer and has a unique graphic interface.

UCSD Signaling Gateway—http://www.signaling-gateway.org/molecule/

- The UCSD signaling gateway molecule pages provide essential information on thousands of proteins involved in cellular signaling.

SPIKE—http://www.cs.tau.ac.il/~spike/

- "SPIKE is a database of highly curated human signaling pathways with an associated interactive software tool."
- Incorporates information from other repositories in the curation process.

BIGG—http://bigg.ucsd.edu/

- "BiGG is a knowledgebase of Biochemically, Genetically and Genomically structured genome-scale metabolic network reconstructions."

HumanNet—http://www.functionalnet.org/humannet/

- "A probabilistic functional gene network of 18,714 validated protein-encoding genes of *Homo sapiens* (by NCBI March 2007), constructed by a modified Bayesian integration of 21 types of 'omics' data from multiple organisms, with each data type weighted according to how well it links genes that are known to function together in *H. sapiens*."

Ingenuity—http://www.ingenuity.com/products/ipa

- Large proprietary database of molecular interactions integrated with analysis tools.

Thomson Reuters Metabase—http://thomsonreuters.com/en/products-services/pharma-life-sciences/pharmaceutical-research/metabase.html

- Large, proprietary "manually curated database of mammalian biology and medicinal chemistry data."

Pathway Studio—http://www.elsevier.com/solutions/pathway-studio

- Large proprietary database integrated with analysis tools.

3.2.4 Related Biological Repositories

BioModels—https://www.ebi.ac.uk/biomodels-main/

- BioModels Database is a "repository of computational models of biological processes." Models described from literature are manually curated and enriched with cross-references.

The Cell Collective—http://thecellcollective.org

- Virtual cell models for simulations.
- Related to NDEx in that they also support a "crowdsourcing" strategy.

BioCarta—http://www.genecarta.com

Curated pathway diagrams. Not a network resource—only diagrams and gene lists are available and no computable connectivity.

3.2.5 Selected Examples of Network-Oriented Analysis

GeneMania—http://www.genemania.org/

- "GeneMANIA finds other genes that are related to a set of input genes, using a very large set of functional association data."

Network Portal—http://networks.systemsbiology.net/

- "Provides analysis and visualization tools for selected gene regulatory networks to aid researchers in biological discovery and hypothesis development."

DAVID—https://david.ncifcrf.gov/

- Gene set analysis enrichment scoring includes pathways.

MSigDB—http://www.broadinstitute.org/gsea/msigdb/index.jsp

- Gene set enrichment analysis.

GenomeSpace—http://www.genomespace.org

- "GenomeSpace is a cloud-based interoperability framework to support integrative genomics analysis through an easy-to-use Web interface."
- Integration includes network-oriented tools.

Cytoscape—http://www.cytoscape.org/

- "An open source software platform for *visualizing* molecular interaction networks and biological pathways and *integrating* these networks with annotations, gene expression profiles and other state data."

3.3 Technical Description

The following section will provide a short technical overview of the NDEx platform and its main components. For more detailed information on any of the following topics, please visit the documentation page on the NDEx website at www.ndexbio.org.

3.3.1 Server and API

The central component of the NDEx framework is the NDEx server, a database application accessed by a web-based **RE**lational **S**tate **T**ransfer **A**pplication **P**rogramming **I**nterface (**REST API**) [11], enabling its use from a wide variety of programming environments. Client libraries to facilitate the use of the API are available for Java, Python, and R. The network exchange format (CX) used in the NDEx API provides a semantic-neutral framework accommodating networks of many types while normalizing the treatment of controlled vocabularies, supporting references, provenance history, and property annotations of elements. The REST interface to the public NDEx server is deployed at http://public.ndexbio.org/rest.

3.3.2 Web User Interface

The main NDEx Web UI is a web application that enables users to search and browse networks, establish accounts, create or join groups, and upload, export, download, and administer networks. It provides simple visualization of small networks and query results from large networks and enables the user to review the provenance history of the network. The user can also edit network metadata, such as title or description. It is important to note that the Web UI relies only on the published NDEx API for access to an NDEx server, so it serves as an example for others to develop large NDEx-based applications. The Web UI for the public NDEx server is deployed at http://public.ndexbio.org.

3.3.3 CX Interchange Format

NDEx v1.3 supports network I/O operations where the network (or subnetwork in a query) is encoded in CX v1.0, the Cytoscape Cyberinfrastructure network interchange format. CX has been developed in conjunction with the Cytoscape group to be the future standard for network interchange by Cytoscape, NDEx, and applications and services in the Cytoscape Cyberinfrastructure (previous versions of NDEx used a JSON network format that will be supported for back compatibility but which is now deprecated).

The purpose of CX is to provide a format in which networks can be transmitted between diverse services (Fig. 1). It is designed for flexibility, modularity, and extensibility and as a message payload in common REST protocols. It enables applications to standardize on core aspects of networks, to coordinate on more specific standards within CX, and to ignore or omit irrelevant aspects. *It is not intended* as an optimized format for storage or for specific functionality in applications.

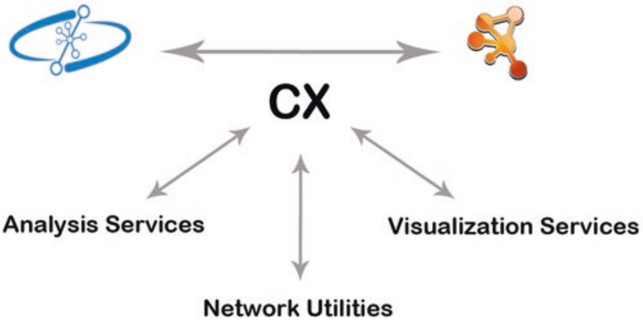


Fig. 1 The CX Interchange Format and its relationships with NDEx and Cytoscape in the context of the Cytoscape Cyberinfrastructure

The context in which CX will transmit networks is the Cytoscape Cyberinfrastructure (CI), a service-oriented architecture (SOA) implemented using REST protocols and primarily intended to facilitate computation with biological networks of many types. Within the CI framework, services may be created to store, query, visualize, lay out, and transform networks. These services may then be orchestrated by applications that implement bioinformatic analyses, ranging from custom user interfaces to interactively scripted applications in environments like IPython Notebook and to managed workflows in environments such as Taverna or Galaxy. An application such as the Cytoscape desktop application can participate both as a client of services and as a service for other applications.

Because of the wide diversity of network formats used in biology, a critical aspect of CX is that it provides straightforward strategies for lossless encoding of potentially any network. At the most basic level, this means that CX imposes very few restrictions: graphs can be cyclic or acyclic and edges are implicitly directed, but formats can choose annotation schemes to override this.

Semantically complex formats such as OWL, BioPAX, OpenBEL, SGML, or SBGN can be supported while at the same time enabling the expression of simple networks without undue overhead. CX does not, itself, make any commitment to a single “correct” model of biology or graphic markup scheme.

Finally, CX also addresses the critical requirement of enabling exchange of large networks while maintaining a low burden on both sending and receiving applications. Particularly, it enables a sending application to begin streaming a network without having the entire network in memory and allows the receiving application to begin processing a network before having received the entire network. Thus, the design of CX enables applications to reduce memory requirements for both the sender and receiver while also lowering delivery latencies.

3.3.4 Data Model

The CX network data model enables the transmission of networks with diverse semantics, uploaded from files in a variety of source formats, including SIF, XGMML, XBEL, and BioPAX3. Expressing these diverse network formats in CX provides users and application developers with consistent handling of nodes, edges, namespaces and identifiers, citations, properties associated with nodes and edges, and network provenance history. The CX data model does not, however, standardize the representation of biology in the networks that it stores. The meaning of the relationships indicated by edges or the classes indicated by the attributes of nodes in a network may conform to a rich standard such as BioPAX3 or OpenBEL, or they may have ad hoc meanings unique to the particular network. NDEx provides a common storage medium and access protocol for CX, facilitating the use of diverse networks by applications but not limiting the semantics that they may express.

The intent in the design of the CX network data model and in the NDEx utilities for loading specific network formats is to fully preserve the information content of networks: a network file in a given format imported to NDEx via CX should be equivalent (though not necessarily identical) to a network file output in a subsequent export operation using that format. The example in Fig. 2 shows one edge represented in the CX network data model.

Although NDEx provides both API and user interfaces to upload files in common formats (XGMML, XBEL, SIF, BioPAX3), the API also provides methods to create and query networks in CX, enabling researchers and developers to create and use networks with arbitrary semantics while still taking advantage of the common infrastructure supported by NDEx (Fig. 3). For example, researchers might experiment with novel representations of RNA–RNA and RNA–DNA interactions using CX aspects for controlled vocabularies, citations, or terminology definition by functional composition. When stored in NDEx, the resulting networks would benefit from NDEx-enabled applications that provide common functions such as basic visualization, indexing for search, or sharing and annotation. Specialized, modular applications (including ad hoc scripts) can then be constructed using the NDEx API to perform analyses and visualization that depend on the novel representation choices. This pattern of use is intended to foster experimentation with representations with rapid, straightforward sharing and discussion of the representational strategies and analytic consequences.

3.3.5 Network Search and Query

An Apache Solr (http://lucene.apache.org/) service running in a separate process provides the primary NDEx text search engine by indexing text strings to elements in the CX data model, including networks, users, groups, nodes, edges, citations, and supports.

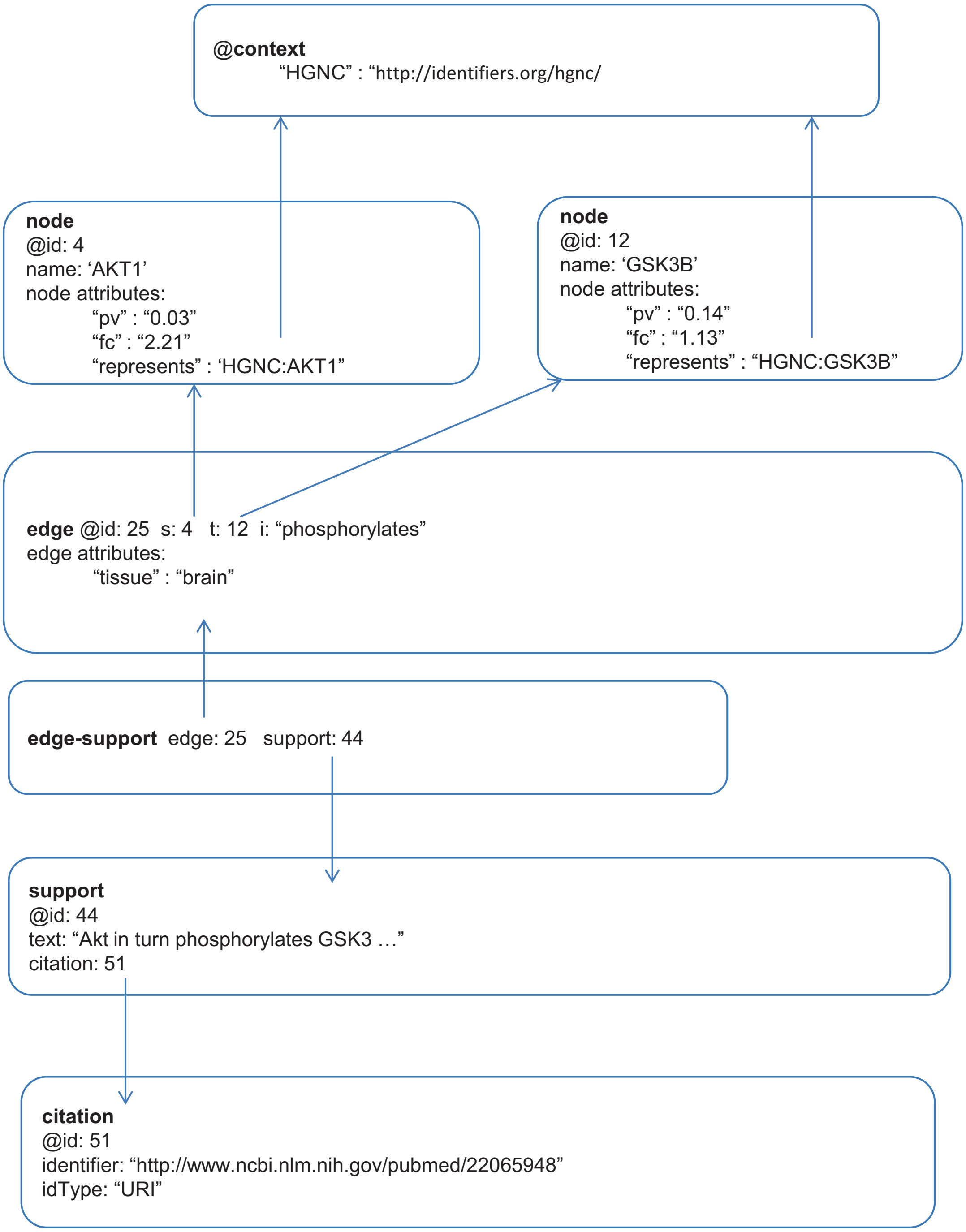

@context
"HGNC" : "http://identifiers.org/hgnc/
node
@id: 4
name: 'AKT1'
node attributes:
"pv" : "0.03"
"fc" : "2.21"
"represents" : 'HGNC:AKT1"
node
@id: 12
name: 'GSK3B'
node attributes:
"pv" : "0.14"
"fc" : "1.13"
"represents" : "HGNC:GSK3B"
edge @id: 25 s: 4 t: 12 i: "phosphorylates"
edge attributes:
"tissue" : "brain"
edge-support edge: 25 support: 44
support
@id: 44
text: "Akt in turn phosphorylates GSK3 …"
citation: 51
citation
@id: 51
identifier: "http://www.ncbi.nlm.nih.gov/pubmed/22065948"
idType: "URI"

Fig. 2 Example of one network edge represented in the CX network data model. Each box in the diagram is a network aspect element, labeled with its aspect name and element id. Edge 25 connects nodes 4 and 12 by the s (source) and t (target) relationships. The meaning of edge 25 is indicated by its "interaction" attribute, "phosphorylates." Nodes 4 and 12 are associated with terms "HGNC:AKT1" and "HGNC:GSK3B" using the "represents" attribute. These terms are linked by their prefixes to the HGNC-controlled vocabulary at a specific reference URI via the @context network aspect, indicating that the terms are standard human gene symbols. Both nodes have user-defined attributes "fc" and "pv" associated with them to record differential expression data mapped onto the network. Edge 25 has a user-defined property "tissue": "brain" used by the authors to indicate the tissue context. Edge 25 is also annotated with an edge-support aspect element, justifying the relationship with evidence text from support 44. Support 44 is in turn derived from citation 51, the article from which the edge was curated

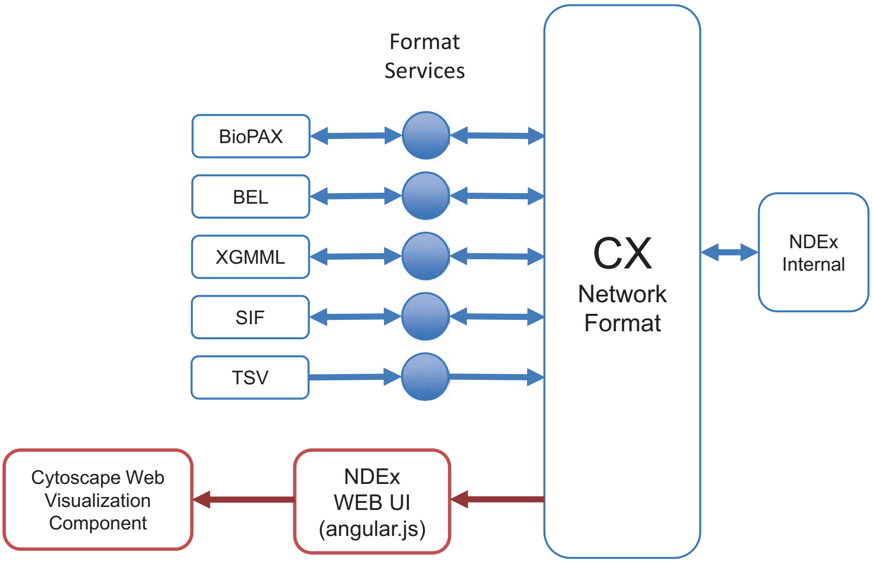


Fig. 3 NDEx v1.3 network format transformations

The Solr indexing enables search for networks by selected attributes. It also enables the identification of network nodes by their names and by controlled vocabulary terms, including terms in specialized CX aspects such as function terms.

The Solr service can run on the same machine instance as the NDEx server, a different machine instance, or even on a cluster. This flexibility provides an easy way to scale this aspect of NDEx, distributing the memory and processor requirements beyond a single machine. The NDEx architecture is explained in Fig. 4.

Because Solr is a standard, widely used search application, the configuration of the indexing of NDEx elements by text will involve less custom software and will be easier to modify, customize, and maintain. It is anticipated that this strategy will also result in improved performance, benefiting from Solr optimizations and future enhancements. The greater ease of customization will facilitate experimentation in indexing and improve application-specific optimizations. Any NDEx server, including the NDEx public server, can potentially extend the default indexing scheme.

The design for a default schema for attribute indexing using the Solr service is described in Table 1. The intent is to limit the default schema to a core set of network attributes that are justified/required in common use cases (such as the NDEx Web UI). The easy customization possible in this design enables a strategy of incremental testing and adoption of further attributes. This attribute schema does not describe the method for indexing networks based on the nodes that they contain. That method incorporates the node name and controlled vocabulary terms associated by CX aspects including nodeAttributes and functionTerms.

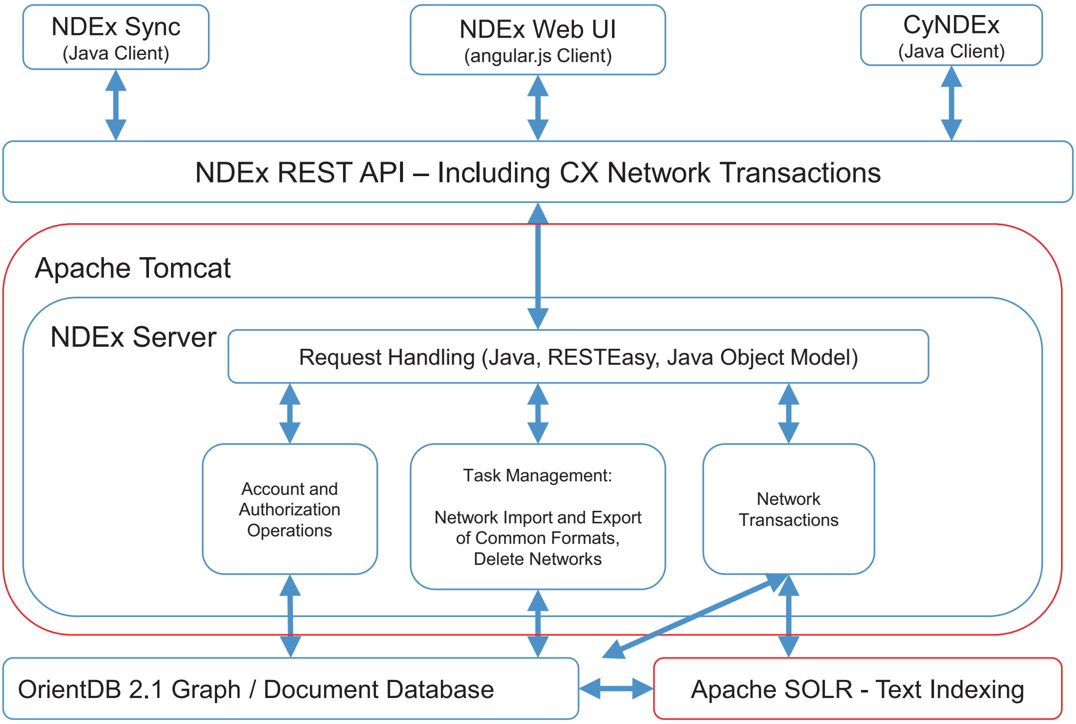

NDEx Sync
(Java Client)
NDEx Web UI
(angular.js Client)
CyNDEx
(Java Client)
NDEx REST API – Including CX Network Transactions
Apache Tomcat
NDEx Server
Request Handling (Java, RESTEasy, Java Object Model)
Account and Authorization Operations
Task Management:
Network Import and Export of Common Formats, Delete Networks
Network Transactions
OrientDB 2.1 Graph / Document Database
Apache SOLR - Text Indexing

Fig. 4 NDEx v1.3 architecture

3.3.6 Usage Metrics

Although NDEx is a relatively young project, it has managed to gain a lot of momentum, thanks to its ambitious set of goals, collaborative efforts, and user-centric philosophy. Figure 5 shows some usage metrics in a simple graphical form. The future development and implementation of administrative tools will greatly enhance the spectrum of statistics that can be tracked, thus providing valuable material for analysis and targeted improvement.

4 A New Wave in the Publication Process

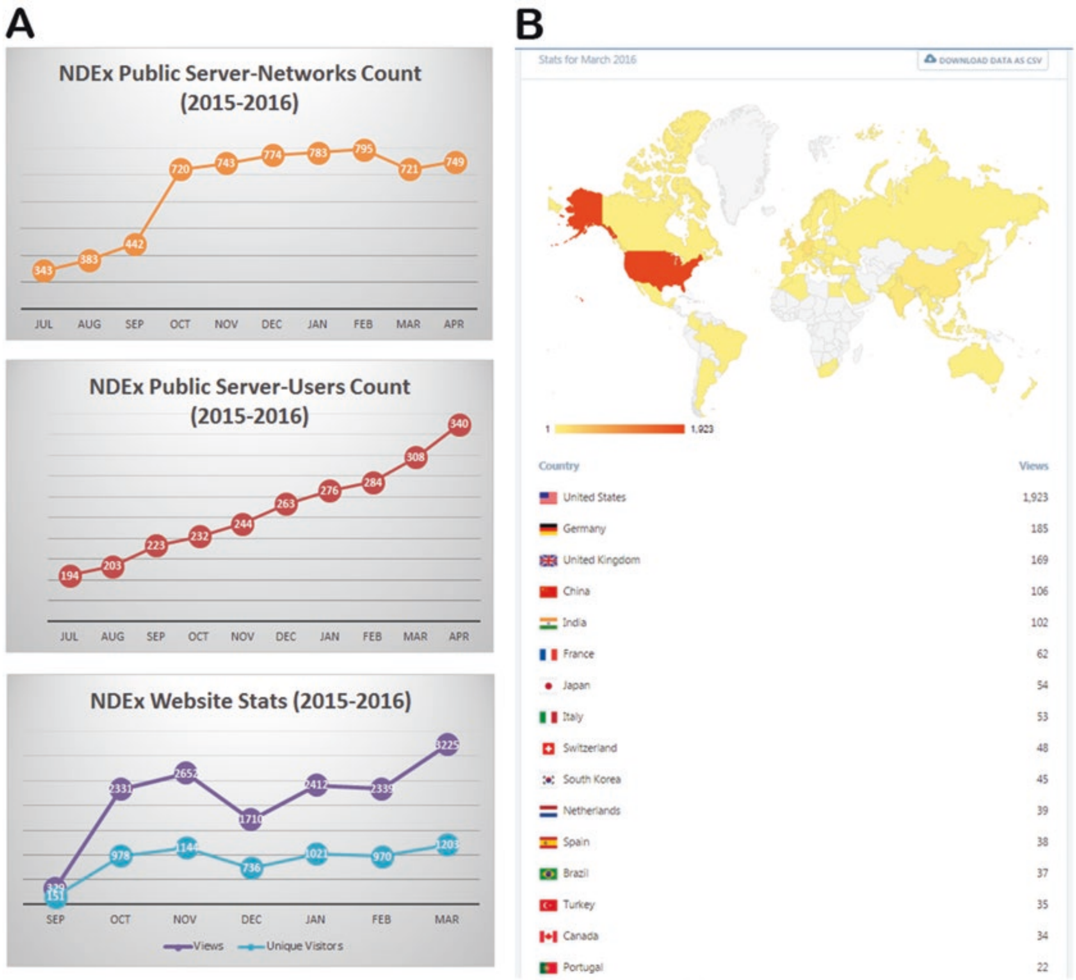
NDEx aims to provide a flexible infrastructure where users can simplify their workflow while promoting collaboration and interaction from the very earliest stages of the research process up until the final communication of results to the public.

4.1 Collaborating Before, During, and After a Publication

As a data commons, NDEx enables scientists and organizations that set up accounts on the NDEx server to upload and save networks and to create communities of users, much like Google+ Circles or LinkedIn groups. Users can manage access to their networks, making them private, public, or shared with selected users and community groups, similar to shared document systems such as Google Docs or Dropbox (Fig. 6). Networks stored in NDEx

Table 1
Default Solr indexing schema. Note that some attributes are highlighted in blue text. These are NDEx internal attributes and are not encoded in the CX networkAttributes aspect. In the case of attributes like "EdgeCount," they are calculated values

Attribute	Description	Type
Name	Name of network	Freetext and string
Description	Description of network	Freetext
Version	Version of network	String
Organism	Description of organisms associated with the network	Freetext and string
Disease	Description of diseases associated with the network	Freetext and string
Tissue	Description of organs, tissue types, cell types, and cell lines associated with the network	Freetext and string
rightsHolder	Describes the holder of the rights to the network—same as dc:rightsHolder	Freetext and string
Rights	Description of the rights asserted by the rightsHolder—same as dc:rights. The indexing of this general field should aggregate not only the actual "rights" attribute but also the more specific rights attributes, such as copyright or license. In that way, network authors do not need to state information redundantly	Freetext and string
Copyright	Copyright asserted by the rightsHolder	Freetext and string
copyrightDate	The date of the copyright asserted by the rightsHolder	Date
licenseType	Identifies the type of license offered by the rightsHolder if a standard form is used	Freetext and string
publicRightsType	Identifies the network as having a public license of one of the following types: unrestricted/academicOnly/notForProfitOnly/researchOnly	String
Contributors	Individuals responsible for creating the content of the network—see dc:contributors	Freetext and string
creationTime	Time at which the network was created	Date
modificationTime	Time at which the network was last modified	Date
edgeCount	Number of edges in the network	Integer
nodeCount	Number of nodes in the network	Integer
Visibility	NDEx network visibility: public, private, discoverable	String
readOnly	NDEx	Boolean
Owner	Name of NDEx account owning the network	String



Country	Views
United States	1,923
Germany	185
United Kingdom	169
China	106
India	102
France	62
Japan	54
Italy	53
Switzerland	48
South Korea	45
Netherlands	39
Spain	38
Brazil	37
Turkey	35
Canada	34
Portugal	22

Fig. 5 Panel (**a**) shows the number of networks (*top*), registered users (*middle*), and numbers of unique visitors and views (*bottom*) for the NDEx public server starting from the NDEx v1.2 official release in July 2015. Panel (**b**) shows the geographical distribution of views in the month of March 2016

can be examined by reviewers during the publication process as live, interactive elements rather than static figures or complex data tables. As an example, the Ideker Laboratory at UC San Diego has recently started using NDEx to control access to networks referenced in submitted publications, thus simplifying and accelerating the review process.

To support publication of networks as data, it must be possible to unambiguously specify the identity of the network and trust that the content of a published network will remain constant. NDEx assigns every network a globally unique ID, a 128-bit UUID that distinguishes it from all others. A UUID reference to a network therefore refers to a specific data artifact, as maintained by its

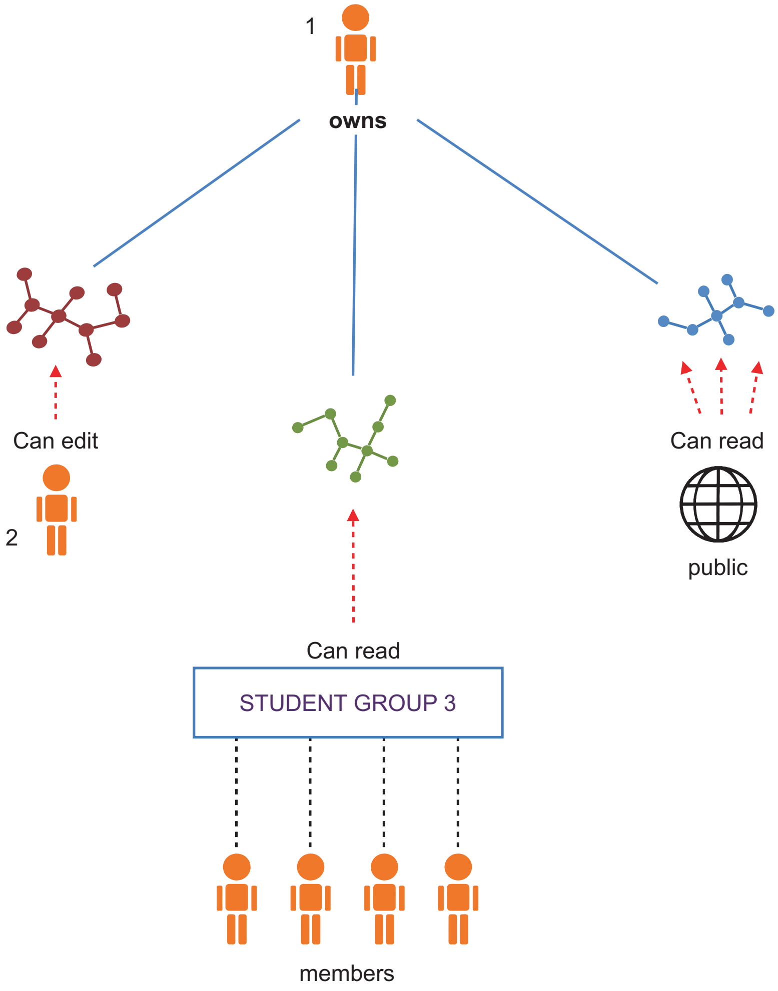


Fig. 6 Example of access control for networks in NDEx. User 1 owns the *red*, *green*, and *blue* networks. She makes the *blue* network public, allowing read access to anyone, including anonymous users; she shares the educational *green* network with the members of Student Group 3 for teaching purposes; finally, she shares the *red* experimental network directly with User 2 for research and collaboration

NDEx server. In addition, the owner of an NDEx network can set its status to be read only, preventing further edits. These features enable networks to be reliable, consistent references, suitable as inputs to further research.

Once published, NDEx seeks to facilitate the reuse of biological networks created by scientists as inputs to further experiment and analysis, providing structures that enable scientific reproducibility and new opportunities for attribution and citation. Given an analysis that uses information from public networks obtained from NDEx, it should be straightforward to reference those specific networks and their authors. Moreover, any researcher wishing to validate the analysis should be able to access the exact networks used or, at a minimum, should know the identity and description of those networks. Finally, when a network is the result of a workflow in which other networks were inputs, it should be possible to access the history of events and sources for easy reproduction of results. At this stage, it becomes important to know how and when a network was created and which inputs and algorithms would be required to reproduce it. NDEx addresses these needs by including the "provenance history" with each network. The provenance history captures the workflow leading to the current network by describing prior events, networks, and other resources and grows as networks are created, modified, used, or copied. The provenance history is described in detail in the next paragraph and Fig. 7.

Finally, organizations that create and publish network content can use NDEx as a channel for distribution: networks from the NCI Pathway Interaction Database [12], the Pathway Commons [13], and the OpenBEL Consortium [14] are few examples of those available in NDEx. As an example, run a "user search" using the keyword "database": NDEx will return all the user accounts of organizations maintaining publicly accessible network databases.

4.2 Network Provenance History

A network can represent assertions of biological relationships that are the results of experimental, analytic, or curation processes. Networks may in turn serve as inputs to further processes of analysis and model creation. If the workflow and dependencies on information sources are clearly documented, researchers may better understand the meaning of the relationships in the network and are better empowered if they wish to reproduce the analyses leading to the network. To achieve these goals, networks stored in NDEx include a provenance history aspect that can be accessed and managed via the NDEx API (Fig. 7).

For example, a network might be derived by an algorithm which finds subnetworks based on experimental data mapped to entities in a reference network; in this case the application performing the analysis should record the analysis event in the provenance history of the output network, including references or descriptions of the algorithm used, the input experimental data, and a description of the input reference network.

For robustness, the provenance history stores descriptions of "ancestor" networks and other information sources, not just links to those resources. This preserves the utility of the provenance history in situations in which some or all of the input information sources are unavailable or have been modified since they were used in the workflow. Researchers (or algorithms) can inspect the provenance history of the current network to address questions about the status of all of the inputs to the workflow. The NDEx network provenance history is similar in intent to synapse analytical provenance [15].

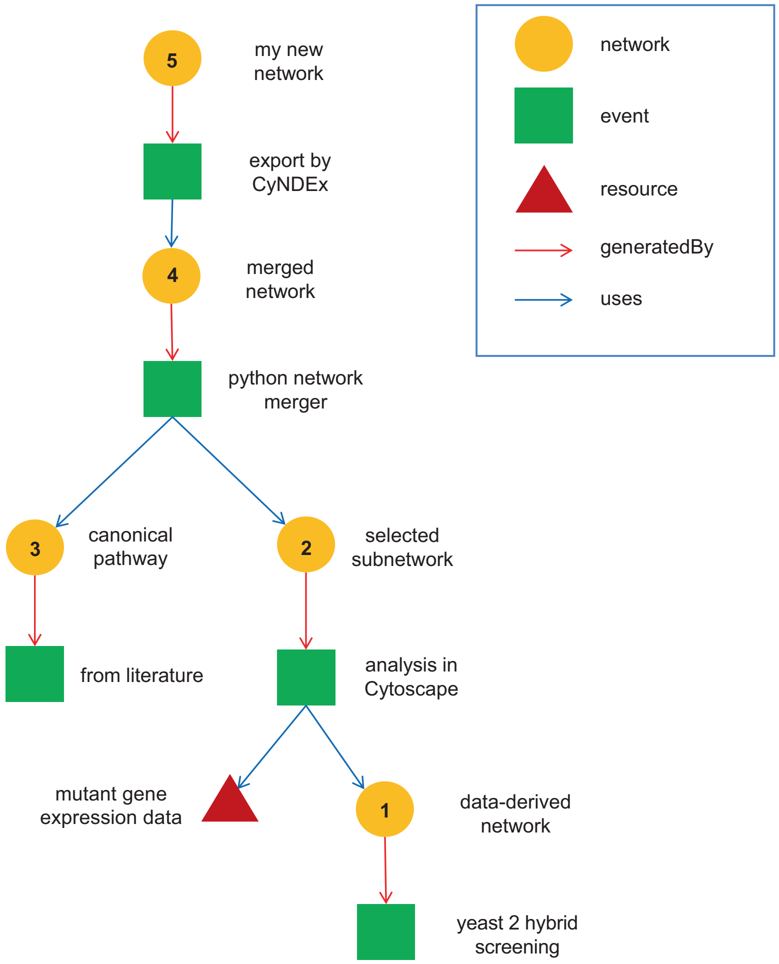


Fig. 7 A provenance history is a tree structure containing ProvenanceEntity and ProvenanceEvent objects. It is serialized as a JSON structure by the NDEx API. The root of the tree structure is a ProvenanceEntity object representing the current state of the network. Each ProvenanceEntity may have a single ProvenanceEvent object that represents the immediately prior event that produced the ProvenanceEntity, which is, in turn, linked to a network of ProvenanceEvent and ProvenanceEntity objects representing the workflow history that produced the current state of the network. The provenance history records significant events as networks are copied, modified, or created, incorporating snapshots of information about "ancestor" networks

The provenance history can be used to infer network equivalence, whether a given network stored in NDEx has the same content as another network or an external resource. This is valuable since in the general case, computing equivalence by algorithm may be computationally expensive or could require network format-specific knowledge. Any operation that modifies the network, including changes to visibility or provenance, also changes the last modification date of the network.

Changes to network membership, what users have access to a particular network, do not modify the network itself and so do not change either the modification date or provenance history.

The standard fields in ProvenanceEntity and ProvenanceEvent objects correspond to relationships defined in the PROV-Ontology (PROV-O) [16]. Other property-value pairs can annotate these objects to provide more information about the entities and events. Any ad hoc pair of strings can be added as a property-value pair, and the properties used may be idiosyncratic to the recorded events and entities. However, the use of properties defined in the Dublin Core (DC) [17] metadata annotations and the Provenance, Authoring and Versioning (PAV) ontology [18] is preferred when applicable.

It is important to note the difference in the use of these ontologies in an NDEx provenance structure and the original intent. A ProvenanceEntity is a description of the referenced object, not the object itself. Therefore, a property such as dc:title that is asserted for a ProvenanceEntity refers to the original entity that the ProvenanceEntity represents. The provenance history references ancestor networks and other data sources but can also include self-contained descriptions of those objects that capture their state at the time they were used.

In a copy operation, an application/utility creates a new network (the target) that encodes the same content as an existing network (the source). In the resulting target provenance history, the root ProvenanceEntity represents the target, and the copy operation is represented as a ProvenanceEvent of type COPY in which the output is the root entity and the input is a ProvenanceEntity representing the source. The ProvenanceEntity representing the source and all of its prior entities and events are copied from the provenance history of the source. Information stored in the provenance history about the source is intended to reflect the state of the source at the time of the copy and should not be updated to reflect subsequent changes in the source. Information about the source stored in the provenance history is thereby preserved, regardless of whether the source is later modified or deleted.

In any case where the source network has the same UUID as the target, the ProvenanceEvent is an edit of some type. Because the event can have both startingAt and endingAt properties, the editing process can span an arbitrary amount of time. The application managing the editing process can therefore control the granularity of the provenance history. For example, an editing application could represent a long sequence of edits in a verbose chain of events and intermediate states, or it could simply keep updating the endingAt time as the edits continued. In both cases, the resulting provenance history would be a valid representation of the workflow, although one would capture greater detail than the other.

In the case where a utility creates a network that has content equivalent or homologous to the source but described in a different identifier system (such as gene ids replaced with corresponding gene symbols), an additional resource describing the identifier mapping is typically involved. In this case, the mapping resource is also an input to the ProvenanceEvent, and it is appropriate to use the property pav:sourceAccessedAt to describe the relationship.

A modification operation in which the information in network A is augmented by information coming from network B, or where a new network is created from both A and B, is called a "merge." This type of modification creates a branched provenance history (Fig. 7), where the ProvenanceEvent for the merge has two inputs, both network A and network B.

5 Applications and Services

Besides its roles as data commons and network publication channel, NDEx provides a flexible, programmatically accessible storage service that promotes modular software development and workflows; in this view, network output from one application can become input to another. NDEx is not meant to perform biological analysis and visualization itself but instead aims to enable the seamless interchange of networks between applications that are specifically designed for analytic purposes.

5.1 Communication Between NDEx Servers: The NDEx Sync Utility

While the public NDEx website is a resource to discover, access, share, or publish networks, there are situations in which it is better to deploy a separate copy of NDEx. For example, an NDEx server can be installed behind a firewall to handle cases where strong security is required, enabling storage of proprietary networks developed for the health science industry or those that incorporate patient information subject to privacy standards (e.g., the Health Insurance Portability and Accountability Act (HIPAA)). A private NDEx can also be deployed on local servers or on a scientist's desktop computer for applications that store very large networks or perform frequent, expensive transactions (Fig. 8).

The information on a private NDEx server will, however, typically be most useful in conjunction with shared and public content. We have therefore created the NDEx Sync, a command-line utility that enables users to synchronize networks between different NDEx server instances, supporting cases where researchers wish to maintain local copies of public reference networks. NDEx Sync can be executed as periodic task, checking specified accounts and networks for changes on a "source" NDEx and then creating or updating networks on a "target" NDEx.

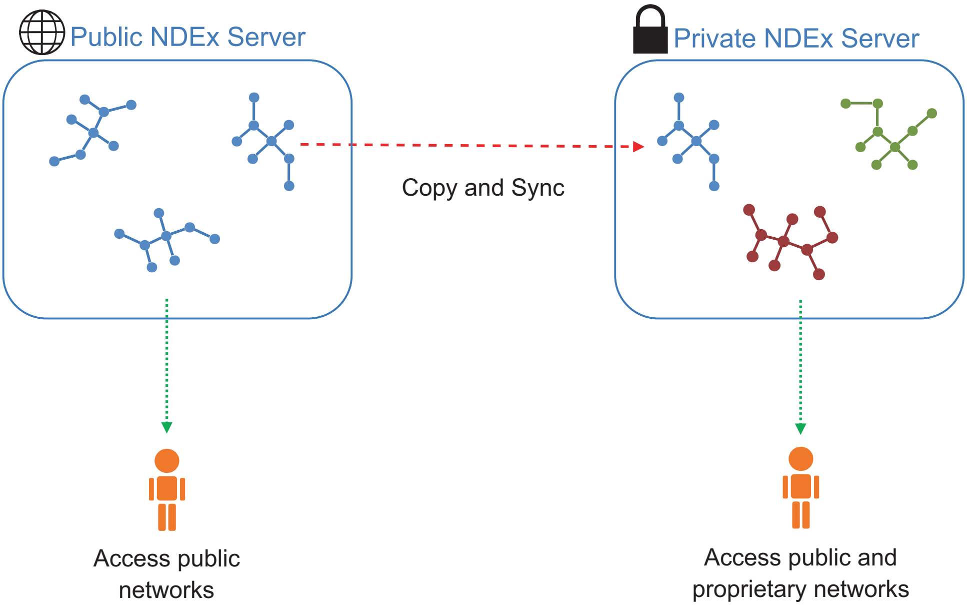


Fig. 8 Communication between NDEx servers. NDEx Sync can be used to copy and synchronize different NDEx servers, thus allowing access to both public and proprietary networks

5.1.1 How It Works

NDEx Sync is like a file-mirroring utility, but with an important difference: the copied networks are not exact duplicates of the source networks. Copied networks are assigned new UUIDs: every network stored in an NDEx server has a *globally unique* identifier and can be referenced by that identifier at its host NDEx. NDEx Sync updates (or creates, if necessary) the network's provenance history, adding a "provenance event" that documents the copy process. The copied networks are therefore documented as *distinct entities*, copied at a specific time from a uniquely identified source. The provenance history provides a structure to document the events leading to the current state of a network. As already discussed, applications using NDEx are not required to maintain the provenance history for networks that they manipulate; however, it is strongly encouraged as a standard practice and will be supported by NDEx utilities.

For each *source network* that is selected as a candidate for copying, NDEx Sync examines the provenance history of *each network in the target account* to determine:

- Was this target network copied from the source network?
- Is the target out-of-date (i.e., the target "date last modified" is older than the source "date last modified")?

The default behavior of NDEx Sync is that it will copy the source network to the target account if there is no copy of the source network in the target account or if the only copies present are out-of-date or have been modified.

5.1.2 Copy Plans

The working instructions for the NDEx Sync utility are defined by the user in a document called a "copy plan"; an example of a copy plan is shown in Fig. 9. The NDEx Sync copy plan specifies:

- An account and credentials for the source NDEx
- An account and credentials for the target NDEx
- The criteria to select networks on the source NDEx, which can be (1) a query to find networks matching search text, (2) a query to find networks administered by an account and matching search text, or (3) a list of network UUIDs
- The **updateTargetNetwork** parameter
- The **updateReadOnlyNetwork** parameter

5.1.3 Update of Networks by NDEx Sync

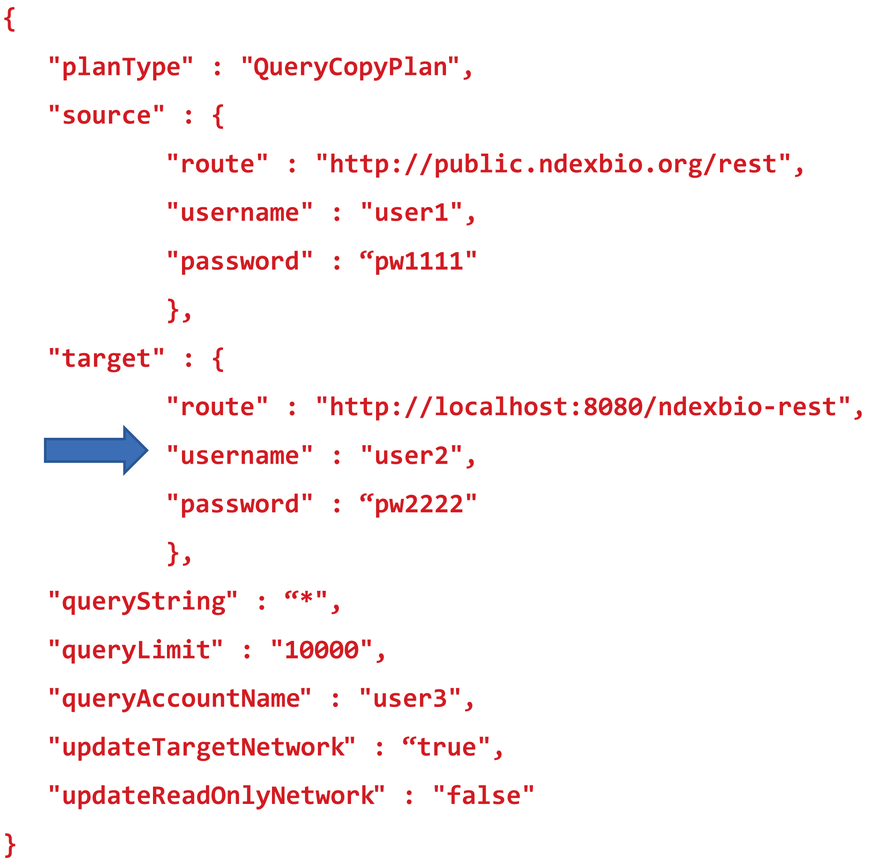
The default behavior of NDEx Sync is conservative, never overwriting or deleting any network in the target directory. This behavior can be overridden by the copy plan parameter **updateTargetNetwork**, specifying that NDEx Sync should update target networks that are identified as unmodified, out-of-date copies of the specified source networks. In an update, the target network keeps its UUID, but its contents are replaced by the contents of the source network, and the provenance history is handled in the same manner as in a default, non-update copy event. The updated network may be accessed by that UUID and any new request will obtain the updated content. Using NDEx Sync to update networks is only appropriate for situations in which the target network is intended as a cache of the source, where users want to obtain the *latest version* of the source content and where they do *not* expect the content of the network to be consistent over time.

5.1.4 Updates of Read-Only Networks

By default, updates will not be performed if the *target* network has read only => true. The **updateReadOnlyNetwork** configuration parameter in a copy plan overrides this behavior. This handles the case in which NDEx Sync is used to maintain a local copy of a remote resource and where the local copy is intended as a read-only, immutable reference. It is important to note that NDEx Sync can only update networks if the account specified by the *username* in the target element in the copy plan (*blue arrow* in Fig. 9) has administration privileges for the networks to be updated.

5.1.5 NDEx Sync and the CX Interchange Format

NDEx Sync uses the CX format for network copy and update operations in which both the source and target NDEx server are v1.3 or later. When a copy or update is performed in CX mode, each of the server transactions is a streamed operation. NDEx Sync does not need to instantiate the entire network in memory; instead, it simply operates by sending the output from the source stream to the input of the target stream. This reduces latency and, most importantly, the memory footprint on each machine involved in the process.



```
{
   "planType" : "QueryCopyPlan",
   "source" : {
            "route" : "http://public.ndexbio.org/rest",
            "username" : "user1",
            "password" : “pw1111"
            },
   "target" : {
            "route" : "http://localhost:8080/ndexbio-rest",
            "username" : "user2",
            "password" : “pw2222"
            },
   "queryString" : “*",
   "queryLimit" : "10000",
   "queryAccountName" : "user3",
   "updateTargetNetwork" : “true",
   "updateReadOnlyNetwork" : "false"
}
```

Fig. 9 Example of NDEx Sync copy plan. With this copy plan, User 2 will copy or update all networks (queryString = *) up to a maximum of 10,000 networks (queryLimit = 10,000) that are available in the User 3 account, from the source NDEx server (http://public.ndexbio.org/rest) to the User 2 account on the target NDEx server (http://localhost:8080/ndexbio-rest). *Blue arrow*: networks in the User 2 account will be updated only if User 2 has administration privileges on those networks

5.1.6 Command-Line Examples

The command is the same as the earlier version of NDEx Sync to run in CX mode:

> bash ndex-copier.sh/users/user12/my-copy-scripts

The "nocx" argument is used to run in the original mode. The original mode is *required* if one or both of the servers being accessed are running NDEx v1.2:

> bash ndex-copier.sh/users/user12/my-copy-scripts nocx

To run NDEx Sync from the MS Windows terminal, use the following commands:

> java -jar ndexbio-sync.jar my-copy-scripts

or

> java -jar ndexbio-sync.jar my-copy-scripts nocx

5.2 Communication Between NDEx and Cytoscape: The CyNDEx App

The Cytoscape desktop application and the wide variety of Cytoscape apps (plug-ins) provide a rich environment for analysis and manipulation of biological networks [19]. Cytoscape can also access NDEx via the REST API, enabling its users to search, query, import, and export networks. Under Cytoscape, a workflow might start by importing a transcriptional regulatory network from NDEx, after which the user could annotate the network with a differential mRNA expression dataset and process it to find subnetworks enriched for genes with significant changes in mRNA expression. The user could then export the subnetworks back to NDEx for review by collaborators or for use as inputs to further analyses.

Communication between NDEx and the Cytoscape desktop application is mediated by the CyNDEx App, and upcoming releases of Cytoscape are expected to incorporate its functionality into the main application, making NDEx networks immediately available to users [8]. The CyNDEx App is an example of an application in which the NDEx Java Client library uses the NDEx API to access NDEx servers.

CyNDEx allows Cytoscape users to connect to a selected NDEx server anonymously or via an existing account, search for networks, and then import a selected network or a subset of a large network as a query result. The imported networks are reduced to nodes, edges, and properties on those nodes and edges, in compliance with the Cytoscape table-oriented data model. Once the work in Cytoscape is done, users can export the modified network to NDEx and make it available to collaborators or the larger scientific community.

5.2.1 CyNDEx and the CX Interchange Format

CyNDEx v1.3 uses the CX format for network transfer from Cytoscape to NDEx. The use of CX enables the storage and retrieval of Cytoscape networks in NDEx without data loss, including the preservation of graphic markup and subnetwork structure. When a network is retrieved (imported) from NDEx, CyNDEx requests all of the CX aspects supported by Cytoscape: the *core* aspects and the *Cytoscape-specific* aspects. In contrast some aspects used in the NDEx Web UI such as citations or functionTerms are not supported by Cytoscape.

If the retrieved network was originally from Cytoscape, the combination of the core aspects and the Cytoscape-specific aspects restores the network in its entirety, bringing layout, graphic markup, and the structure of subnetworks and views. This Cytoscape data structure is displayed as a “network collection,” presenting views of its subnetworks that in some cases can be extremely simple, having only one view. CyNDEx provides two options for saving (exporting) networks to NDEx: either the entire network collection is saved as an NDEx network or a single view may be selected and saved.

If the retrieved (imported) network does not have any Cytoscape aspects, CyNDEx must create default structures in Cytoscape, building a simple network collection with a single view and default graphic style. If either the network collection or the single view is saved (exported) to NDEx, the new structures will also be stored.

Note that if a network with NDEx-specific aspects is retrieved and then saved back to NDEx as a new network, the NDEx-specific aspects *will not* be included in the new network. The only exception to this rule is the provenance history.

5.2.2 Network Update

Using the new NDEx API method **updateCXNetwork**, clients can update selected CX aspects and leave others untouched. CyNDEx can update a network stored in NDEx even though Cytoscape does not support all of the aspects in the original network. For example, a network using the NDEx-specific citation aspect can be imported into Cytoscape without its citations and then have Cartesian coordinates and other properties assigned. Those new properties, along with any other Cytoscape information, can be used to update the network in NDEx without changing the citations stored for that network. As discussed in the CX specification, it is possible to create networks that have inconsistencies between aspects managed by different applications, so applications are responsible for detecting and reconciling inconsistencies when they receive a network.

5.2.3 Provenance History

Although Cytoscape does not support the provenanceHistory CX aspect, CyNDEx preserves the provenance history of the networks it retrieves from NDEx when it creates new networks in NDEx based on these retrieved networks. Like the NDEx server, CyNDEx v1.3 relies on the **CXIO** Java IO library (*see* Materials) created by the Cytoscape team.

5.3 An Example of an NDEx Utility: The Python TSV Loader

Very often, network data consist of complex spreadsheets that can be hard to understand and challenging to interpret without carefully reading the publication in which they appear. The Cytoscape desktop application is capable of loading networks from Microsoft Excel spreadsheets or Tab-separated (TSV) text files; however, users are limited in their freedom to attach custom properties to the different network elements. In addition, there is no easy way to handle the citation information associated with nodes and edges. To overcome these limitations, we have created a versatile and fully customizable **TSV Loader** that takes advantage of the NDEx Python Client library (*see* Materials). The NDEx Python Client v2.0 was released as a PyPi module that can be installed via the PIP installation utility, and a detailed Python Client tutorial is available in the documentation section of the NDEx website at www.ndexbio.org.

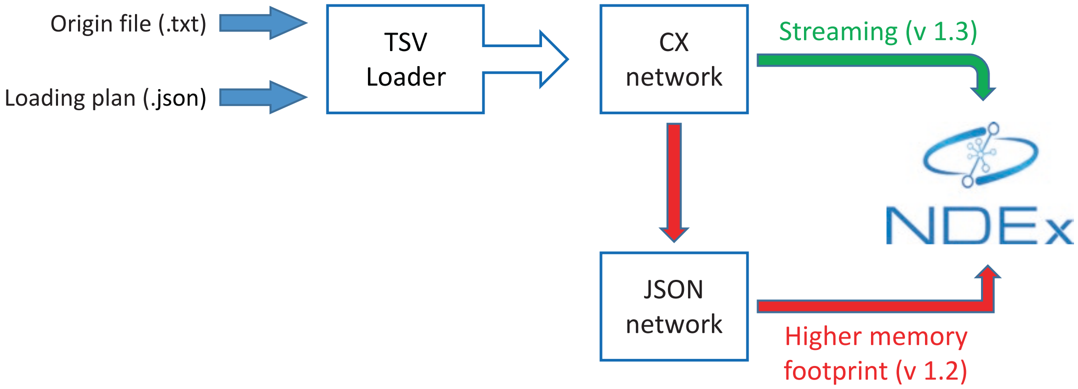


Fig. 10 NDEx TSV Loader. In NDEx v1.2, the TSV Loader creates a CX network, converts it to a JSON network, and then uploads it to NDEx. In v 1.3, the CX network will be directly uploaded to NDEx using the new CX streaming capability

5.3.1 Python TSV Loader

The TSV Loader has been developed to work with NDEx v1.2, the latest officially released version deployed to the NDEx public server at the time this chapter was written. However, we have taken into account the upcoming implementation of CX in NDEX v1.3 and made sure the TSV Loader was flexible enough to adapt to this new standard while being unaffected by the gradual phase out of the older platform.

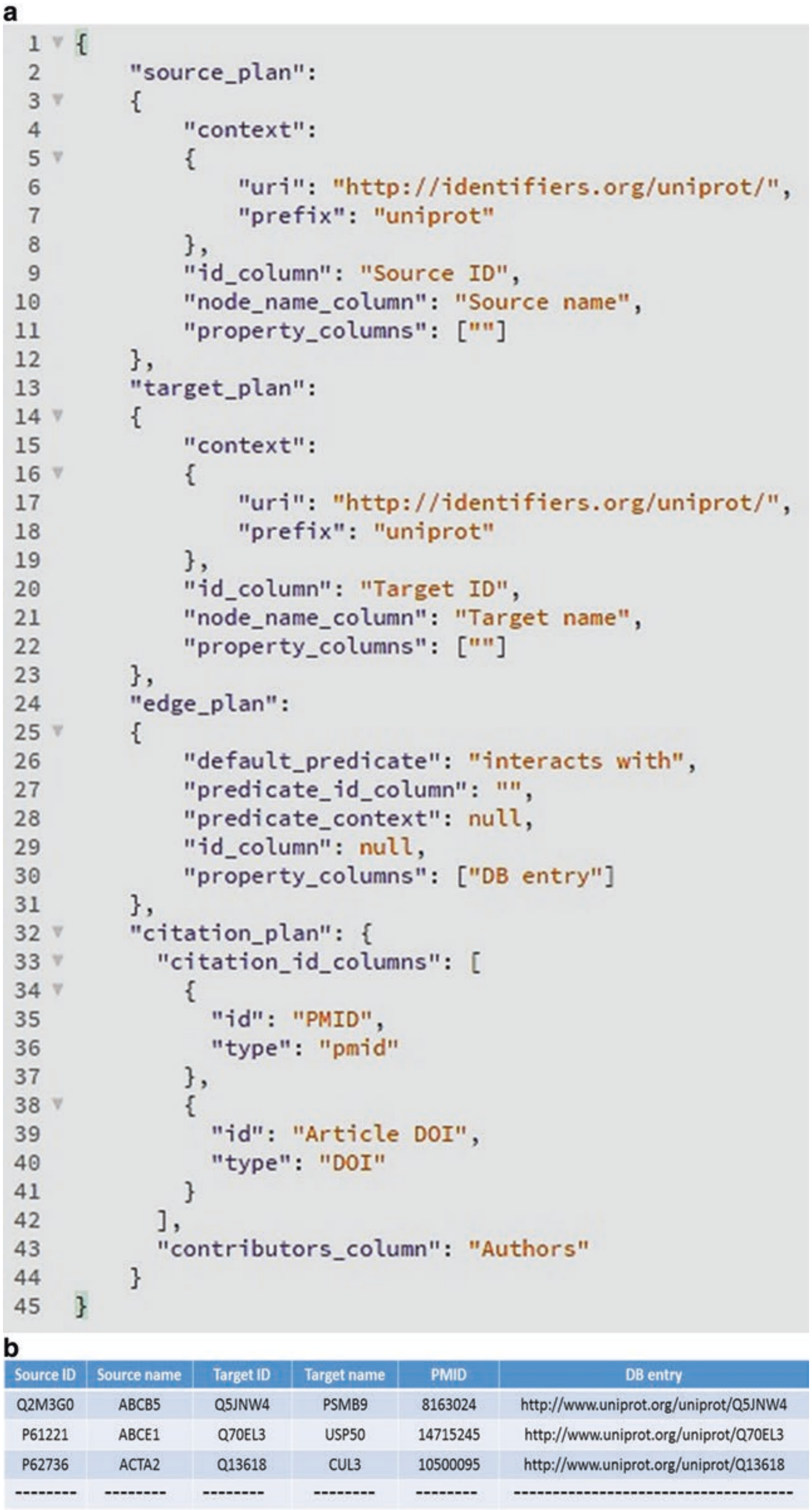
In NDEx v1.2, the TSV Loader creates a CX network, transforms it in a JSON network, and then uploads it to the NDEx server (*red* path in Fig. 10). This requires the *full* network to be accommodated (stored) in the server's memory, which can turn out to be a very expensive operation in cases where multiple users upload large networks at the same time.

In NDEx v1.3, the TSV Loader will create a CX network and upload it directly to the NDEx server using the CX streaming feature (*green* path in Fig. 10). This will dramatically reduce the memory footprint on the server's side and guarantee better performance and efficiency. The next phase of development will wrap the TSV Loader utility in an intuitive UI and deploy it as a stand-alone service.

5.3.2 Origin File and Loading Plans

The TSV Loader is a command-line utility that, similarly to the NDEx Sync, uses JSON plans as templates for parsing networks for upload to NDEx. In addition to the loading plan, the TSV Loader also requires a tab-separated text file (origin file) that contains the actual network data. An example origin file and loading plan are shown in Fig. 11.

The loading plan consists of four main sections: source plan, target plan, edge plan, and citation plan (Fig. 11a). In each section, the user can specify what columns from the origin file (Fig. 11b)



```
{
  "source_plan":
  {
    "context":
    {
      "uri": "http://identifiers.org/uniprot/",
      "prefix": "uniprot"
    },
    "id_column": "Source ID",
    "node_name_column": "Source name",
    "property_columns": [""]
  },
  "target_plan":
  {
    "context":
    {
      "uri": "http://identifiers.org/uniprot/",
      "prefix": "uniprot"
    },
    "id_column": "Target ID",
    "node_name_column": "Target name",
    "property_columns": [""]
  },
  "edge_plan":
  {
    "default_predicate": "interacts with",
    "predicate_id_column": "",
    "predicate_context": null,
    "id_column": null,
    "property_columns": ["DB entry"]
  },
  "citation_plan": {
   "citation_id_columns": [
    {
     "id": "PMID",
     "type": "pmid"
    },
    {
     "id": "Article DOI",
     "type": "DOI"
    }
   ],
   "contributors_column": "Authors"
  }
}
```

Source ID	Source name	Target ID	Target name	PMID	DB entry
Q2M3G0	ABCB5	Q5JNW4	PSMB9	8163024	http://www.uniprot.org/uniprot/Q5JNW4
P61221	ABCE1	Q70EL3	USP50	14715245	http://www.uniprot.org/uniprot/Q70EL3
P62736	ACTA2	Q13618	CUL3	10500095	http://www.uniprot.org/uniprot/Q13618
--------	--------	--------	--------	--------	-----------------------------------

Fig. 11 (**a**) Example of a loading plan for the TSV Loader: users can fully customize the plan including namespaces, IDs, citation, and node/edge properties. (**b**) Sample origin file showing three protein–protein interactions with external links and PubMed IDs; the column names in this origin file are the same used in the loading plan shown in panel (**a**)

should be used to generate the network. The TSV Loader can easily handle citations, whether they are PubMed IDs or DOIs. Users are also free to append as many properties as they like to either nodes (source and/or target), edges, or both. In all those cases where a property is specified by a column in the origin file that contains a full URL ("DB entry" column in Fig. 11b), the TSV Loader will generate a clickable element, thus allowing seamless access to external resources right from within NDEx.

5.4 Outreach and Collaborations

The NDEx Project is carrying on several collaborations to develop and integrate analytic applications and utilities that will be essential elements of the Cytoscape Cyberinfrastructure. The CRAVAT/Mupit tool developed by the Karchin Lab [20] will use NDEx networks as back end for its enrichment analysis, while the Fraternali group at King's College London is developing an analytic application for short loops in protein–protein interaction networks [21] that will use NDEx as source of networks, storage platform, and visualization. Additional collaborations with projects supported by the NCI ITCR program are still at a planning stage and will be starting in Q2 2016.

Since its inception, the NDEx Project has been developed keeping in mind that researchers might not be expert computational biologists; NDEx should be intuitive enough to be used by any researchers, from bioinformaticians to wet lab molecular biologists. In order to engage and educate users, the NDEx Project maintains a strong social media presence including a Twitter page (@NDExProject), a LinkedIn company page (https://www.linkedin.com/company/ndex-project), and a YouTube channel currently hosting only the NDEx Overview video but soon to be the reference point for video tutorials and user guides.

The NDEx Project is involved in the NCI Informatics Technology for Cancer Research Training and Outreach Workgroup (ITCR TOW) and in the J. Craig Venter Institute Network Special Interest Group (JCVI Network SIG) for the purpose of coordinating and improving outreach activities such as workshops, training sessions, and meetings.

In addition, the NDEx Project maintains an informational website (www.home.ndexbio.org) with news, blog, and the **NDEx Documentation**, both for users (manuals and tutorials) and developers (technical literature).

Acknowledgments

The NDEx Project is a joint effort of the Cytoscape Consortium and the Ideker Lab at the UC San Diego School of Medicine.

Primary financial support for The NDEx Project is from the National Cancer Institute (U24 CA-184427), F. Hoffmann-La Roche Ltd., Janssen Research and Development, LLC, and Pfizer, Inc.

NDEx is also supported by the National Resource for Network Biology (P41 GM103504).

References

1. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42:D472–D477
2. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
3. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1998) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34
4. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. BMC Bioinf 9:399
5. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
6. Califano A, Butte AJ, Friend S, Ideker T, Schadt E (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 44:841–847
7. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. In Nature methods (United States). pp. 1108–1115
8. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T (2015) NDEx, the network data exchange. Cell Systems 1(4):302–305
9. CyNetShare (2014) http://cynetshare.ucsd.edu
10. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. PLoS Biol 6:e184
11. Fielding RT, Taylor RN (2002) Principled design of the modern Web architecture. ACM Transactions on Internet Technology (TOIT) 2:115–150
12. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the pathway interaction database. Nucleic Acids Res 37:D674–D679
13. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway commons, a web resource for biological pathway data. Nucleic Acids Res 39:D685–D690
14. OpenBEL (2011) http://www.openbel.org
15. Synapse Analytical Provenance (2015) https://www.synapse.org
16. PROV-O (2013) PROV-O: The PROV Ontology http://www.w3.org/TR/prov-o/
17. DublinCore (2012) Dublin Core Metadata Element Set, Version 1.1 http://www.dublincore.org/documents/dces/
18. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJG, Goble C, Clark T (2013) PAV ontology: provenance, authoring and versioning. J Biomed Semantics 4:37
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
20. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R (2013) CRAVAT: cancer-related analysis of VAriants toolkit. Bioinformatics 29(5):647–648
21. Chung SS, Pandini A, Annibale A, Coolen AC, Thomas NS, Fraternali F (2015) Bridging topological and functional information in protein interaction networks by short loops profiling. Sci Rep 5:8540