

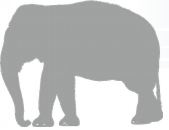
Toward Community-Driven, Shared Literature Annotation Resources

Jin-Dong Kim
Database Center for Life Science
(DBCCLS)

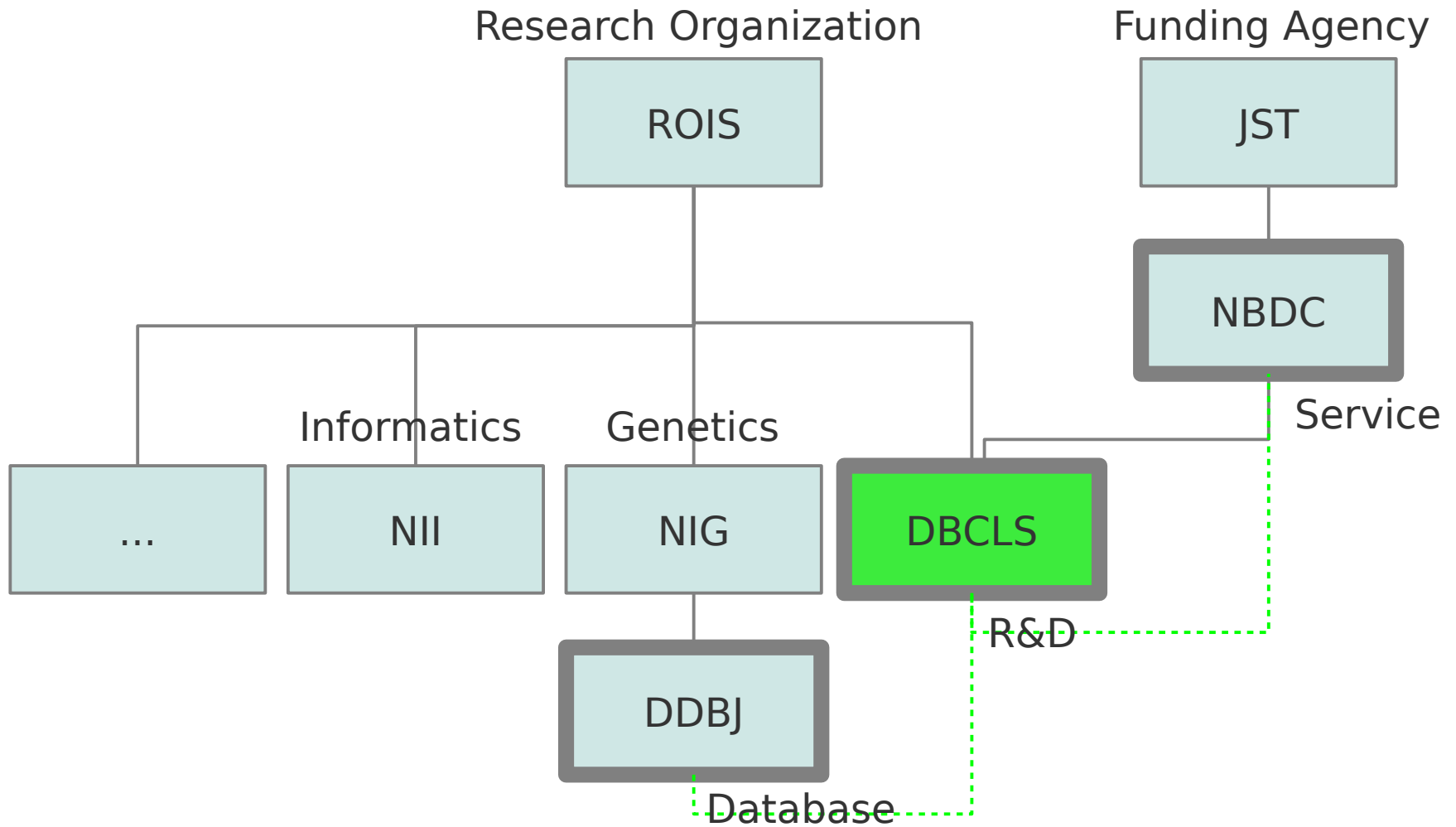


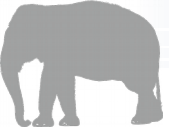
DBCLS

- Database Center for Life Science
 - ✓ A government-funded research center
 - ✓ For integration of databases of life sciences
 - ✓ It annually organizes
 - ➔ BioHackathon series
 - ➔ BLAH series

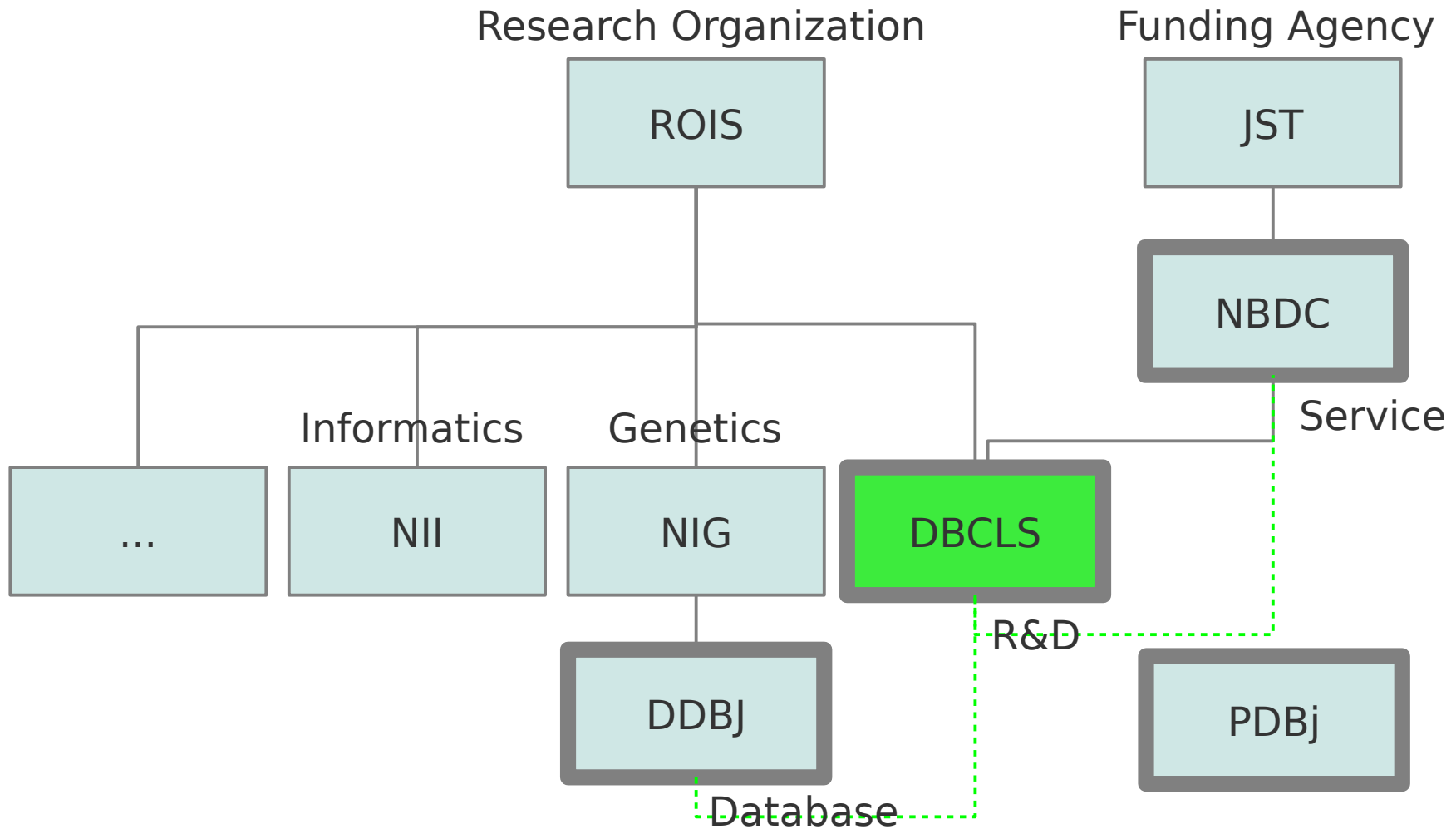


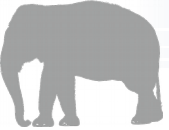
DBCLS



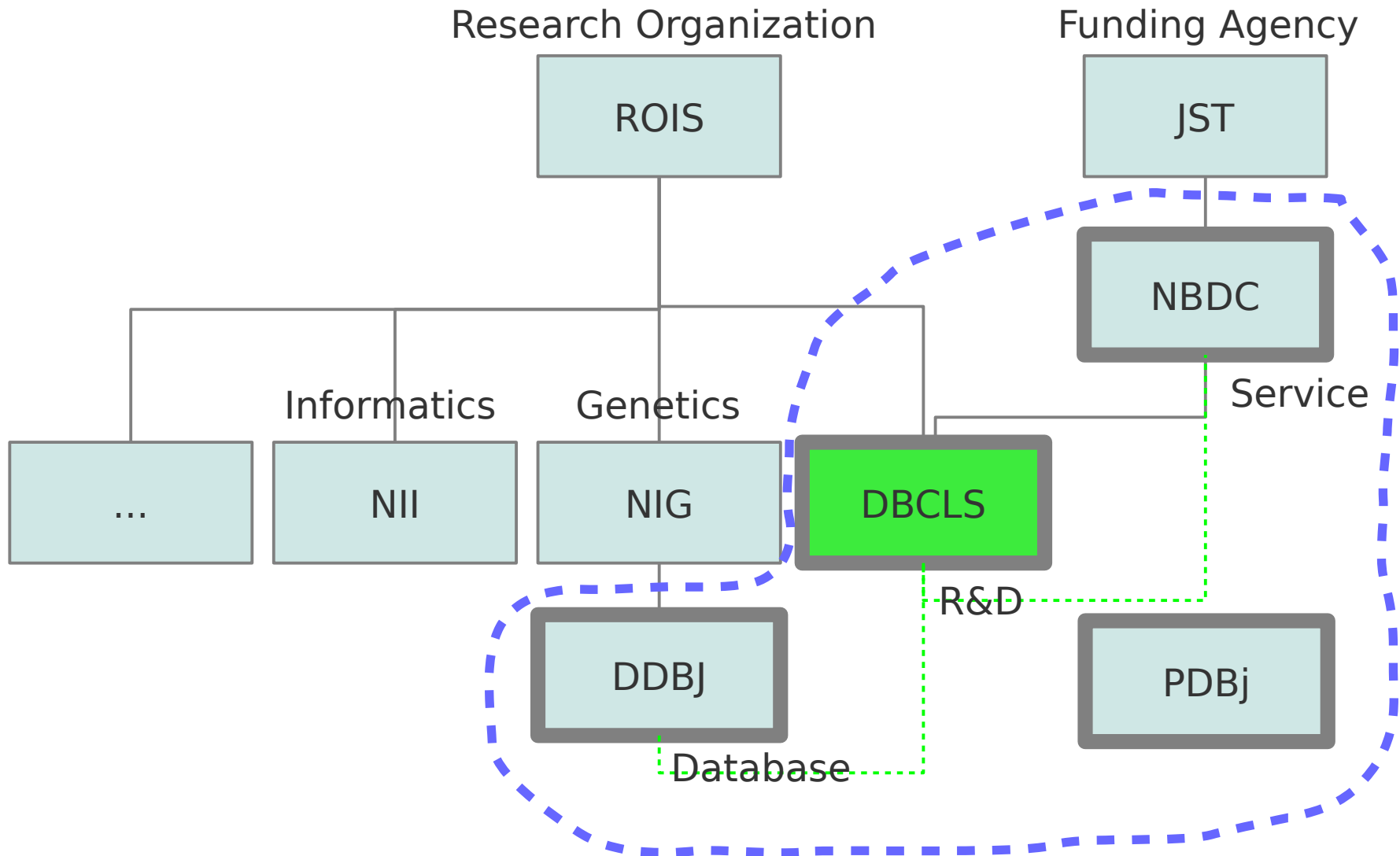


Introduction to DBCLS





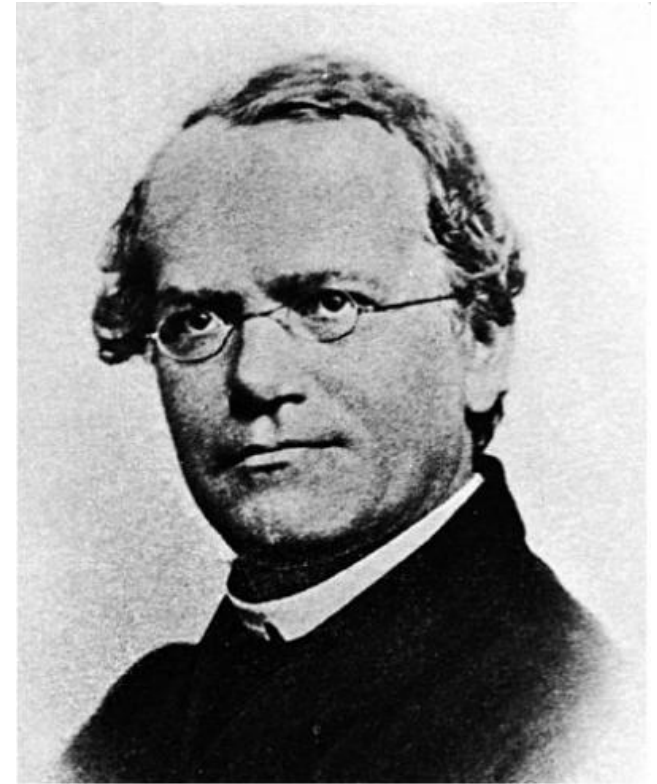
Introduction to DBCLS





“Inheritance involves the passing of discrete units of inheritance”

*Proceedings of the
Natural History Society of
Brünn, 1866*



Gregor Mendel

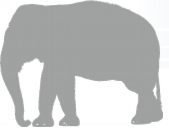


“Inheritance involves the passing of discrete units of inheritance”

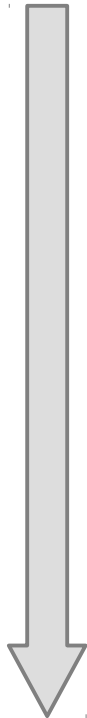
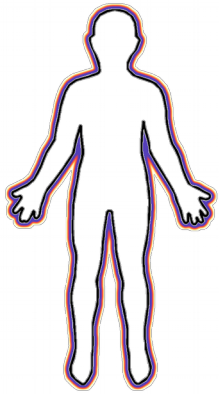
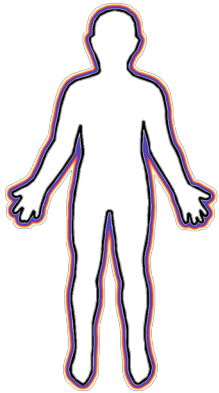
*Proceedings of the
Natural History Society of
Brünn, 1866*



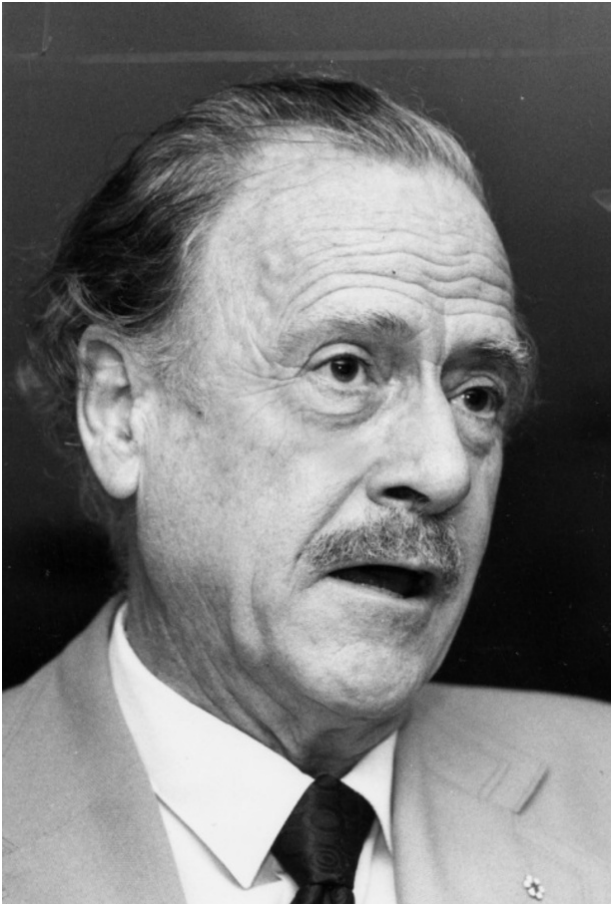
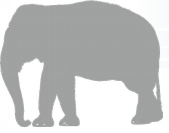
Gregor Mendel



Heredity



genome



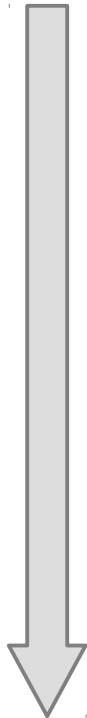
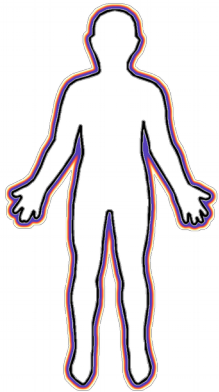
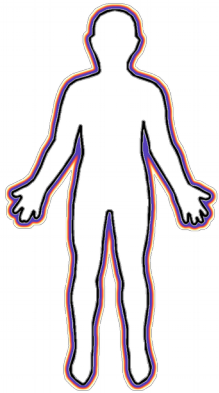
Marshall McLuhan

“all media are extensions of
our human senses, bodies
and minds.”

The Medium Is the Massage,
1967



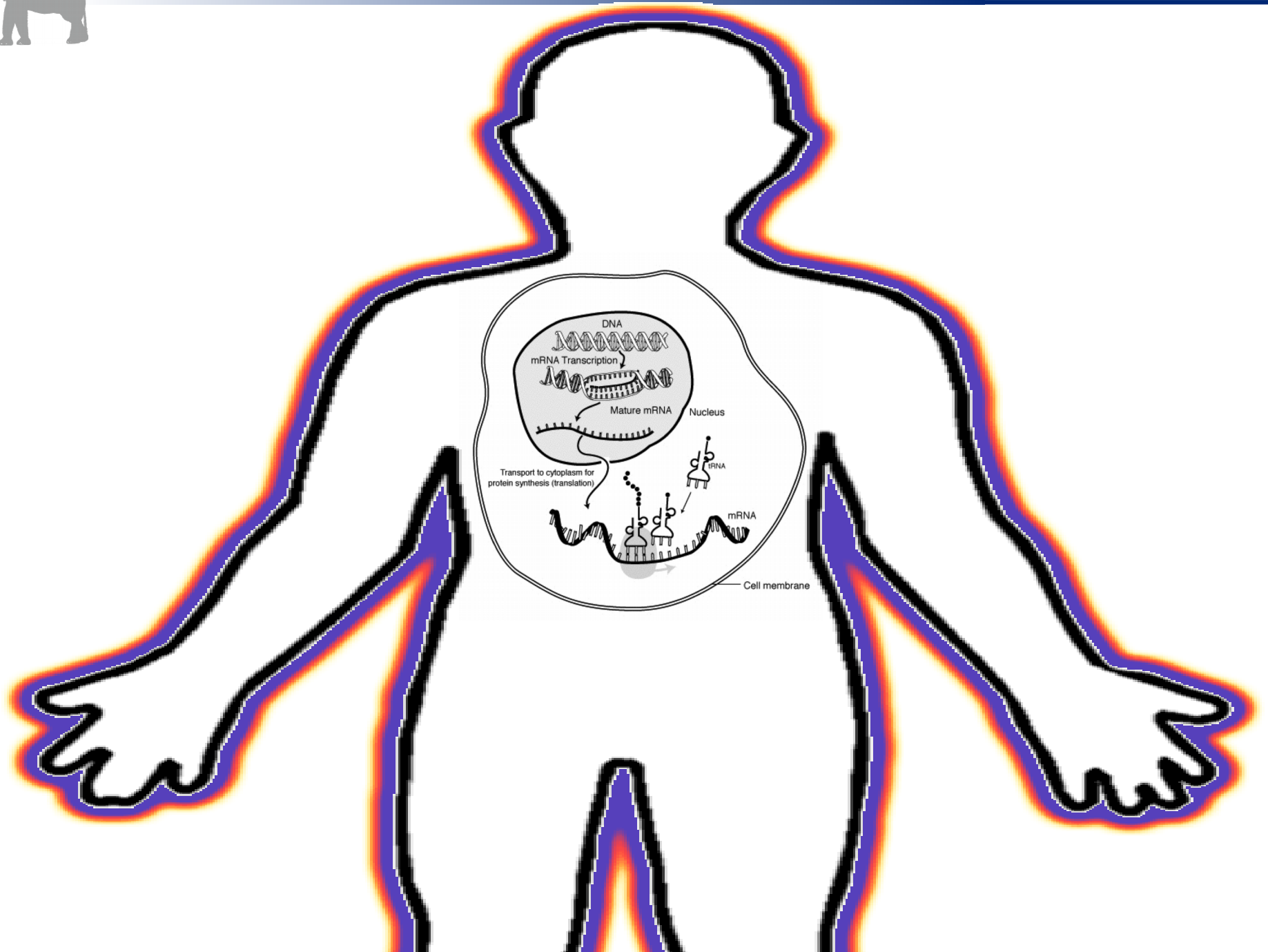
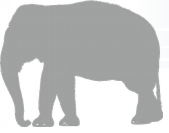
Extension of Heredity

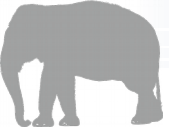


genome

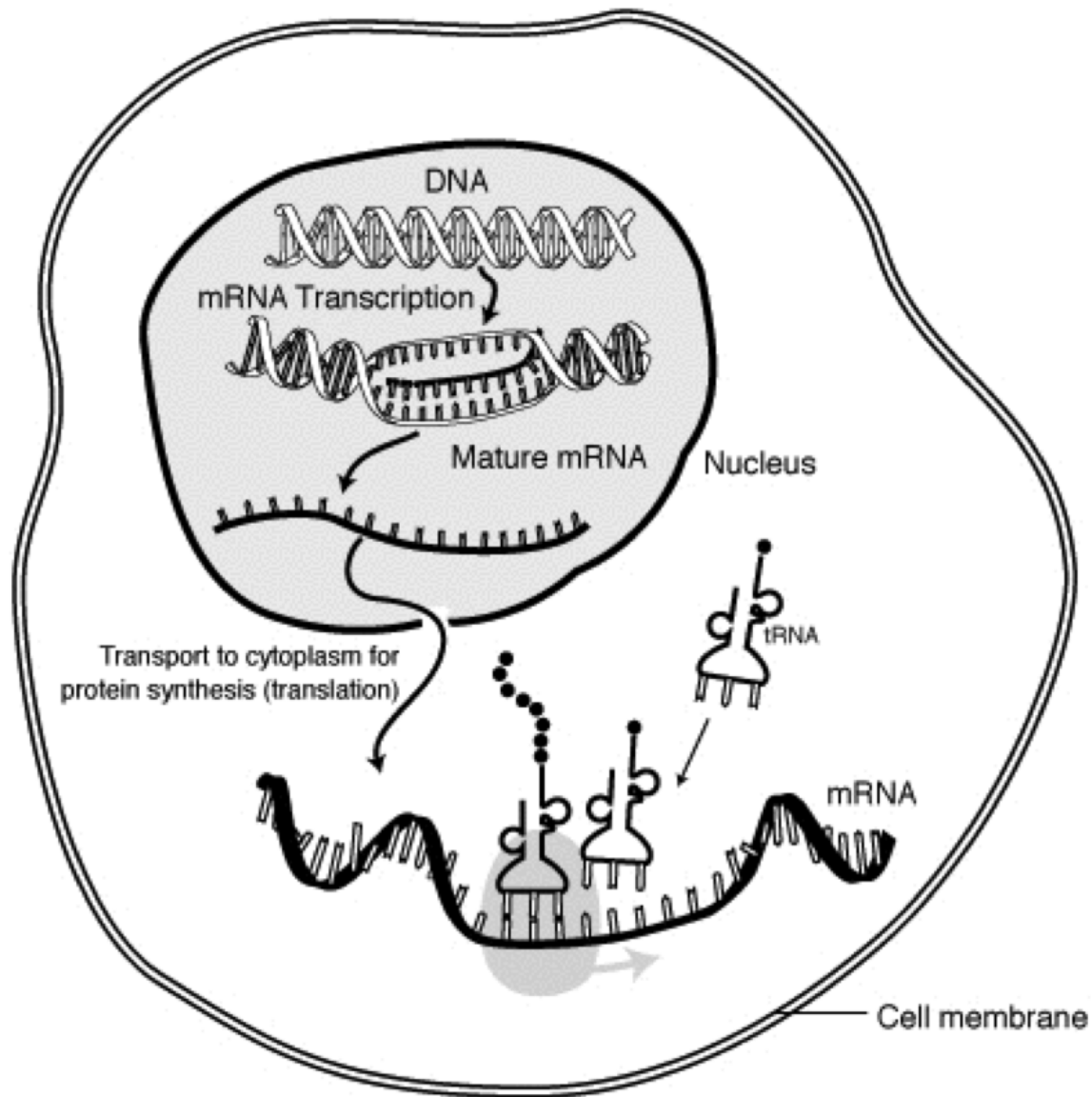


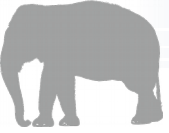
literature





Cell





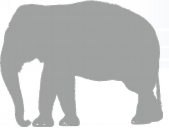
Society



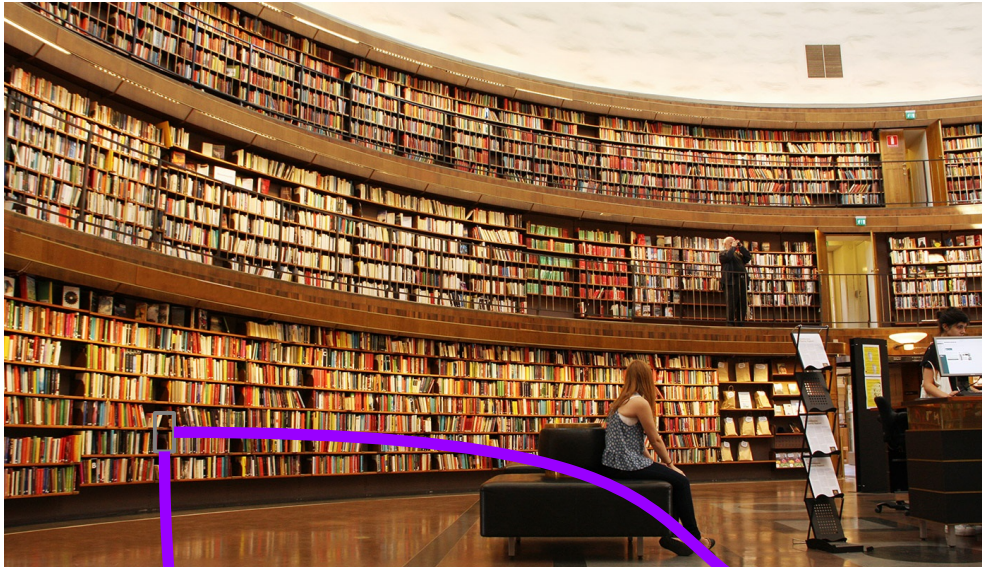
transcribed

translated



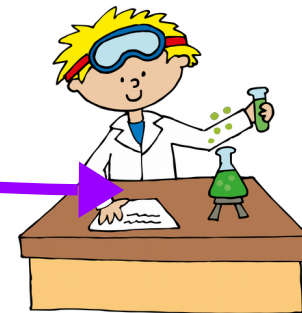


Society



transcribed

translated





Science Literature

- Scholarly articles, Textbooks, ...
- Authoritative information
- Accumulation of scientific knowledge
- Basis for new discoveries
- Repeatedly accessed



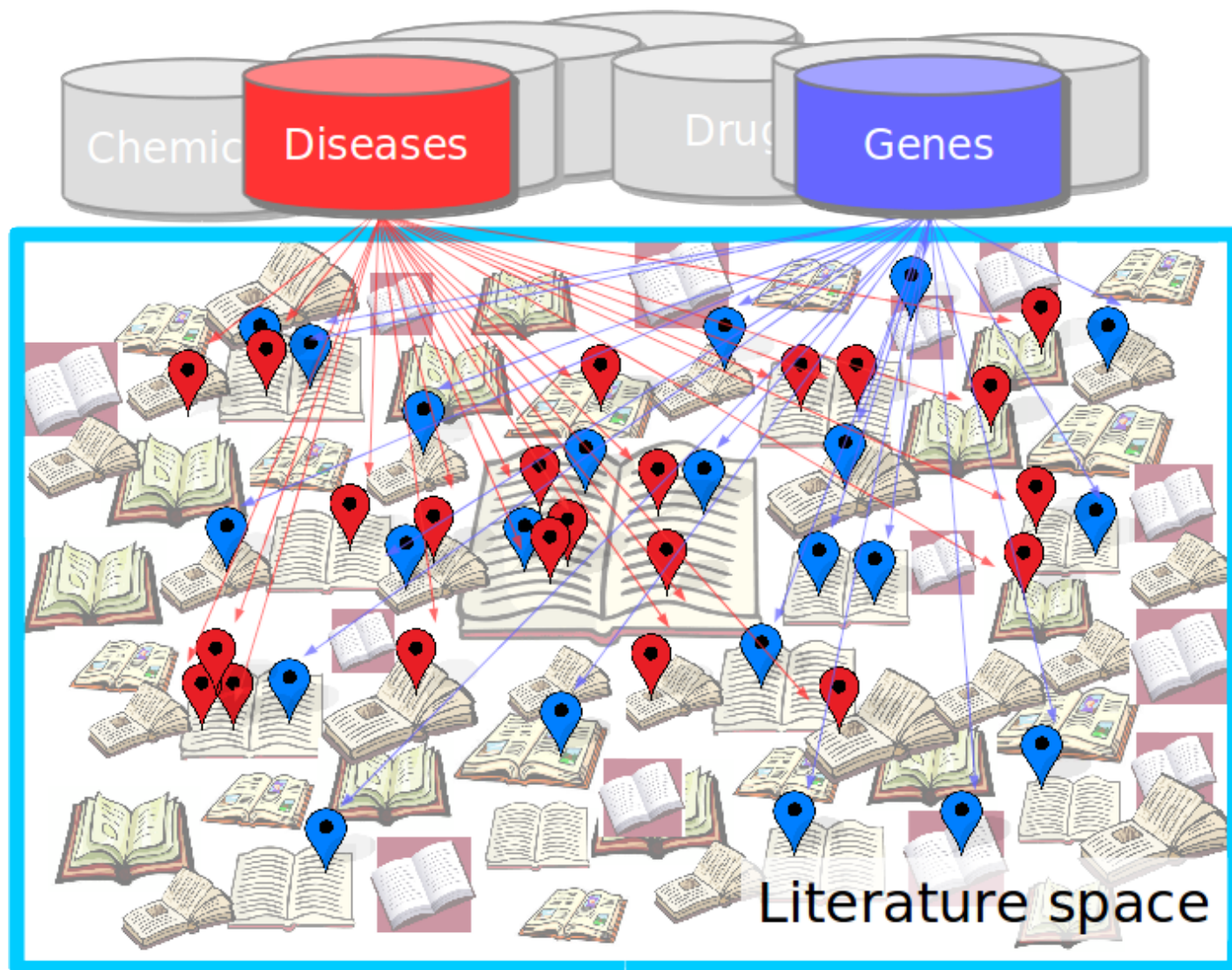


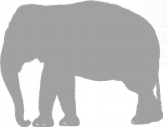
My research is based on ...



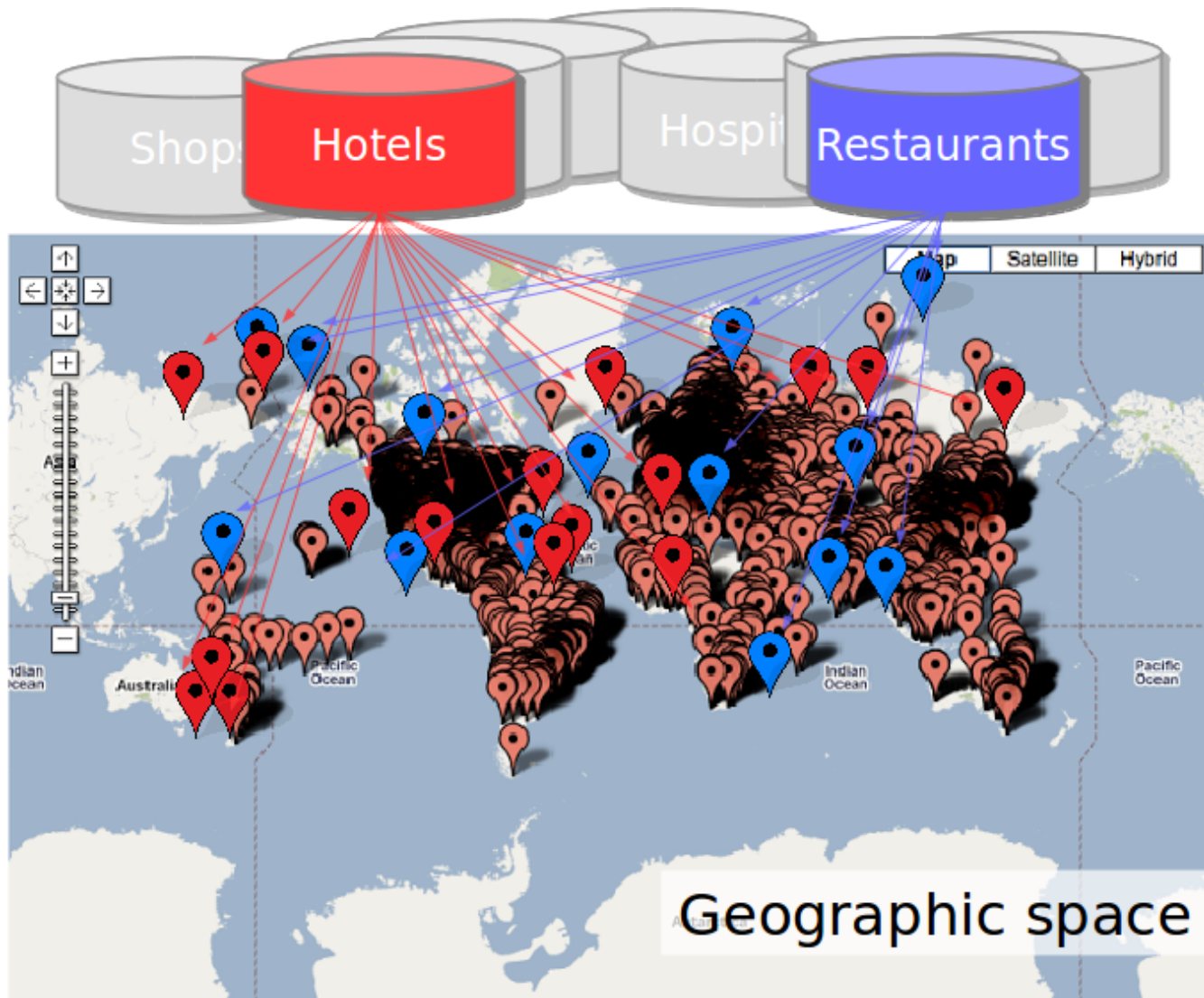


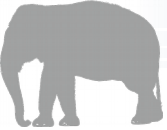
Literature Indexing



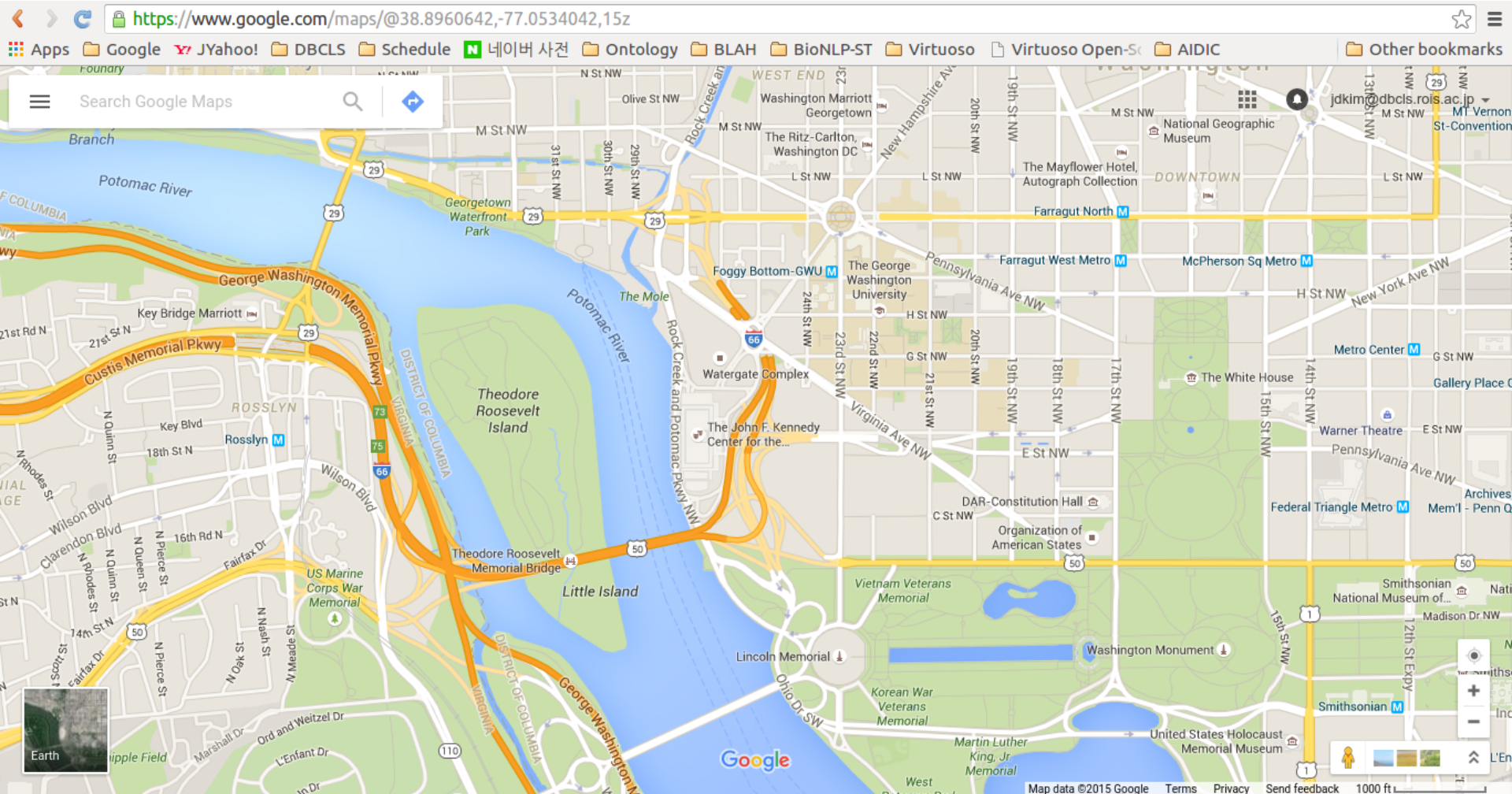


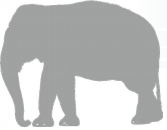
Geospatial Indexing





Google Maps (Washington D.C.)



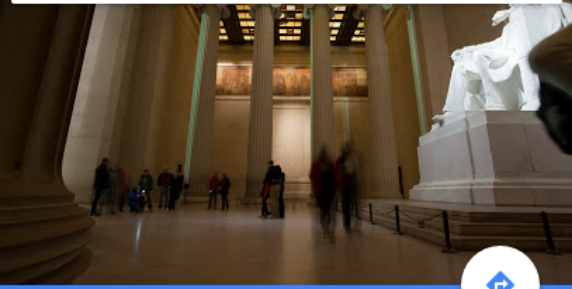


Entities

<https://www.google.com/maps/place/Lincoln+Memorial/@38.8960642,-77.0534042,15z/data=!4m2!3m1!1s0x0000000000000000:0x1049d1c9c95c2eb6>

Apps Google JYahoo! DBCLS Schedule 네이버 사전 Ontology BLAH BioNLP-ST Virtuoso Virtuoso Open-S AIDIC Other bookmarks

Lincoln Memorial



Lincoln Memorial

4.7 ★★★★★ 522 reviews

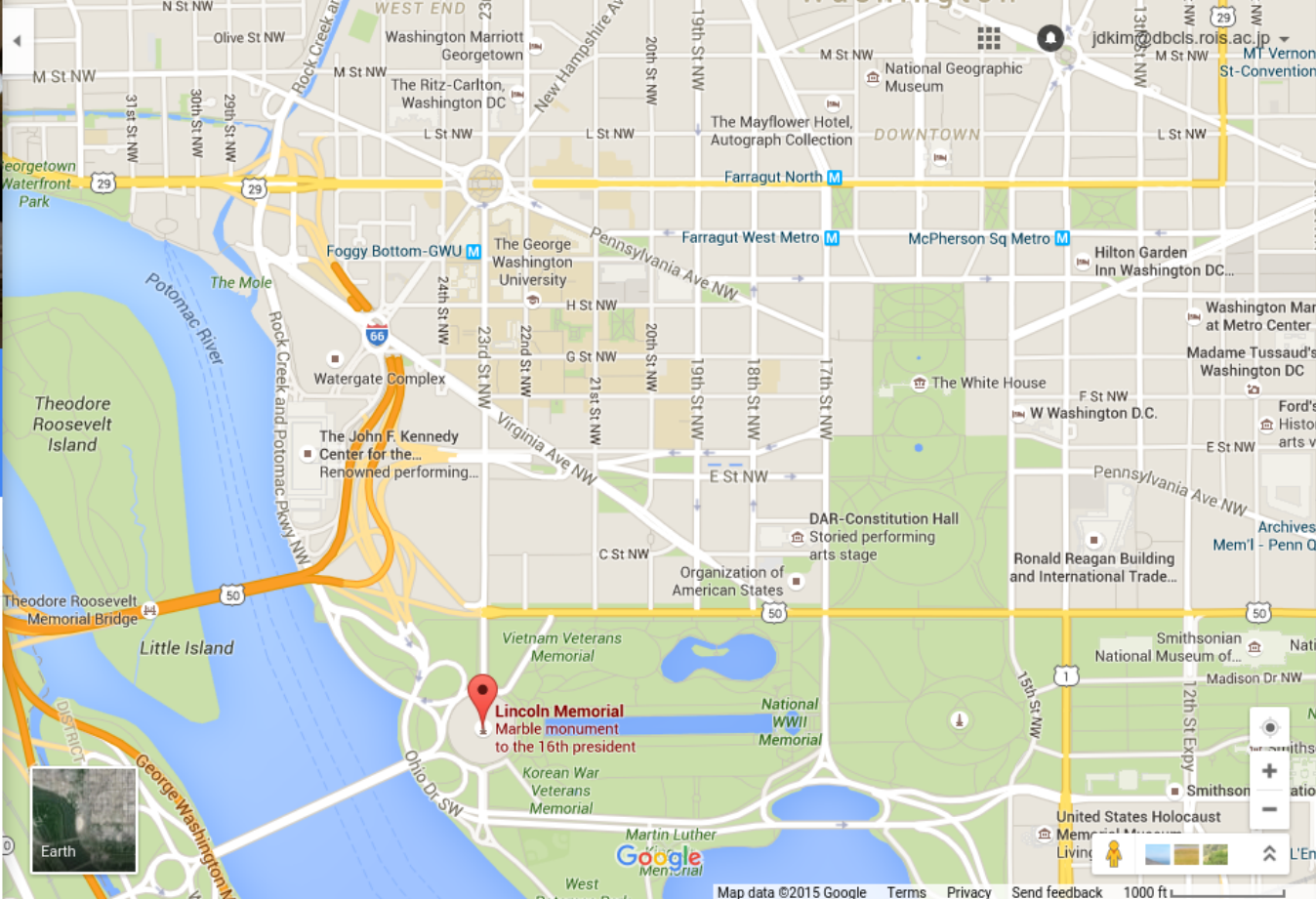
[Monument](#)

Directions

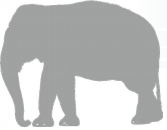
SAVE NEARBY SEND TO DEVICE SHARE

Parthenon-inspired tribute to Abraham Lincoln with a 19-ft. marble statue, murals & reflecting pool. - Google

- 2 Lincoln Memorial Cir NW, Washington, DC 20037
- nps.gov
- (202) 426-6841
- Open now: Open 24 hours
- [Suggest an edit](#)



Map data ©2015 Google Terms Privacy Send feedback 1000 ft



(Geospatial) Pathways

https://www.google.com/maps/dir/Farragut+North/Lincoln+Memorial,+2+Lincoln+Memorial+Cir+NW,+Washington,+DC+20037,+United+States/@38.8960642,-77.0534

Apps Google JYahoo! DBCLS Schedule 네이버 사전 Ontology BLAH BioNLP-ST Virtuoso Virtuoso Open-S AIDIC Other bookmarks

Farragut North Station
Lincoln Memorial, 2 Lincoln Memorial C

OPTIONS

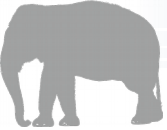
- via 17th St NW 29 min 1.4 miles
DETAILS
- via 18th St NW 29 min 1.4 miles
- via 19th St NW 30 min 1.5 miles

Map data ©2015 Google Terms Privacy Send feedback 1000 ft



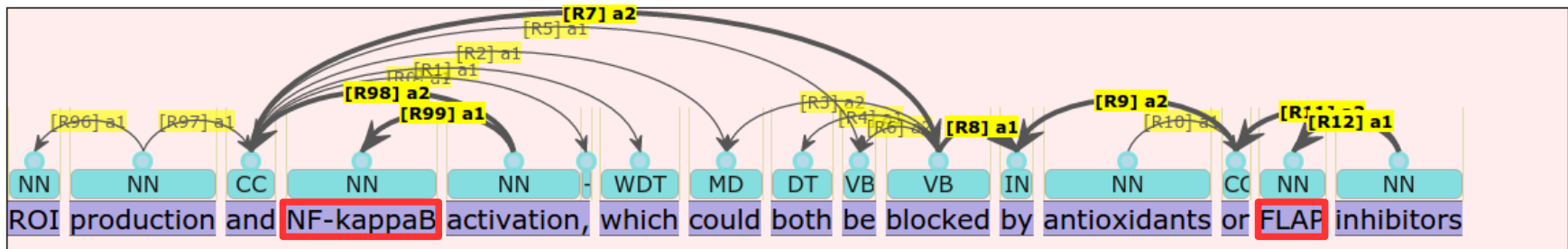
Entities

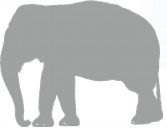
ROI production and **Gene_4790** **NF-kappaB** activation, which could both be blocked by antioxidants or **Gene** **FLAP** inhibitors



(Linguistic) Pathways

Gene_4790
ROI production and NF-kappaB activation, which could both be blocked by antioxidants or FLAP inhibitors

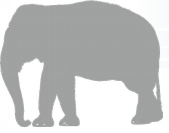




Why people use Google Maps?

- Useful
 - ✓ Contents
- Easy to use
 - ✓ Interface
 - ➔ To access
 - ➔ To exchange
 - ➔ To create
 - ➔ To reuse

Geospatial annotations



Literature annotation

- Do we have good contents?
 - ✓ Many groups are producing annotations.
 - Do we have good ways to access them?
 - ✓ Ann. resources are scattered and isolated.
- ➡ Let's link them to each other, and
Share them, altogether. *BLAH!*



PubAnnotation

- ✓ Is a repository of literature annotation
- ✓ Is based on a scalable storage system
- ✓ Specifically aims at PubMed and PMC
- ✓ Solicits contribution of annotations from the community
- ✓ Solves various problems for sharing the annotations
 - ➔ Alignment
 - ➔ Global addressing system
 - ➔ REST APIs



Make your annotation public, and more useful!

REPOSITORY SEARCH NEWS GUIDES ABOUT

English 日本語 [signup](#) [login](#)

[> top](#)

Documents



PMC	9,044	<input type="text" value="keywords"/>		<input type="text" value="source ID"/>	
PubMed	8,424,657	<input type="text" value="keywords"/>		<input type="text" value="source ID"/>	
FirstAuthor	8	<input type="text" value="keywords"/>		<input type="text" value="source ID"/>	

Projects (146)



With most annotations

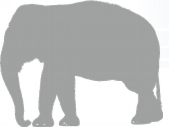
Name		# Ann.	Updated At	Status
PubmedHPO		12,437,742	2016-01-31	Uploading
DisGeNET		3,117,504	2016-01-28	Beta
NEUROSES		2,151,082	2016-02-24	Beta
CRAFT-treebank		844,123	2015-11-19	Beta
bionlp-st-ge-2016-spacy-par...		225,680	2016-05-25	Released
spacy-test		136,597	2016-05-25	Released
FSU-PRGE		59,505	2016-05-17	Released
craft		52,960	2015-10-13	Beta
jnlpba-st-training		51,290	2016-09-09	Released
PennBioIE		23,881	2016-05-17	Released

News

- (09 Mar 2016) Recent access problem
- (25 Feb 2016) System problem on 25/02/2016
- (19 Jan 2016) a new project status, Uploading, added.
- (04 Jan 2016) News service begins.

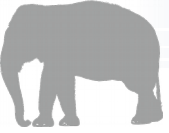
Recently updated

Name		# Ann.	Updated At	Status
jnlpba-st-training		51,290	2016-09-09	Released
jnlpba-st-test		6,005	2016-09-07	Uploading
bionlp-st-seedev-2016-training		109	2016-09-05	Uploading
bionlp-st-bb3-2016-training		1,292	2016-09-05	Released
bionlp-st-cg-2013-training		10,935	2016-08-22	Released
bionlp-st-pc-2013-traing		7,855	2016-08-22	Released
bionlp-st-id-2011-training		5,609	2016-08-22	Released
bionlp-st-epi-2011-training		7,595	2016-08-22	Released
Ab3P-abbreviations		2,342	2016-07-29	Beta
AnEM_full-texts		689	2016-07-27	Uploading



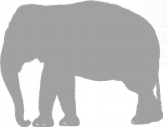
Challenges for Integration

- Format is not standardized
 - ✓ Many proprietary formats
- Texts are changed
 - ➔ PubMed, PMC change texts
 - ➔ Web masters change texts
 - ➔ Annotation projects change texts
- ✓ For
 - ➔ Cleaning
 - ➔ Convenience for annotation
 - Unicode → ASCII



Challenges for Integration

- Format is not standardized
 - ✓ A matter of conversion
- Texts are changed
 - ✓ Breaks stand-off annotation
 - ➔ Character offsets become invalid
 - ✓ Solution
 - ➔ Sequence alignment (BLAST!)



Alignment

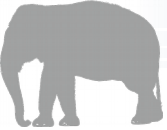
PubAnnotation

GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated that interleukin 4 (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also confirmed in Th2 cells and in T cells of transgenic mice overexpressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF) β -mediated FOXP3 induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. ...

Local Annotation

This inhibitory mechanism was also confirmed in Th2 cells (208-213, Protein transgenic mice over-expressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF)- β -mediated FOXP3 induction.

107-113, Protein



Alignment

PubAnnotation

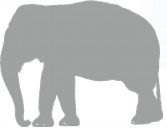
GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated that interleukin 4 (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also confirmed in Th2 cells and in T cells of transgenic mice overexpressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF) β -mediated FOXP3 induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. ...

Local Annotation

Upload & align

This inhibitory mechanism was also confirmed in Th2 cells & 208-213, Protein transgenic mice over-expressing GATA-3 in T cells, which are shown to be deficient in transforming growth factor (TGF)-beta-mediated FOXP3 induction.

107-113, Protein



Alignment

PubAnnotation

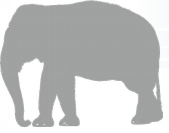
GATA3-Driven Th2 Responses Inhibit TGF-1Induced FOXP3 Expression and the Formation of Regulatory T Cells Transcription factors act in concert to induce lineage commitment towards Th1, Th2, or T regulatory (Treg) cells, and their counter-regulatory mechanisms were shown to be critical for polarization between Th1 and Th2 phenotypes. FOXP3 is an essential transcription factor for natural, thymus-derived (nTreg) and inducible Treg (iTreg) commitment; however, the mechanisms regulating its expression are as yet unknown. We describe a mechanism controlling iTreg polarization, which is overruled by the Th2 differentiation pathway. We demonstrated **838-833, Protein** (IL-4) present at the time of T cell priming inhibits FOXP3. This inhibitory mechanism was also **936-941, Protein** Th2 cells and in T cells of transgenic mice overexpressing **GATA-3** in T cells, which are shown to be deficient in transforming growth factor (TGF) β -mediated **FOXP3** induction. This inhibition is mediated by direct binding of GATA3 to the FOXP3 promoter, which represses its transactivation process. ...

Local Annotation

Upload & align

This inhibitory mechanism was also confirmed in Th2 cells **208-213, Protein** transgenic mice over-expressing **GATA-3** in T cells, which are shown to be deficient in transforming growth factor (TGF)-**beta**-mediated **FOXP3** induction.

107-113, Protein



Alignment

- Jin-Dong Kim, “A generalized LCS algorithm and its application to corpus alignment”, Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP), pp.14-18, 2013
 - ✓ A definite solution to text variant problem
 - ✓ It can align even full paper articles sourced by two different groups.

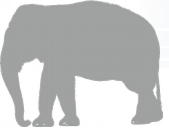


Aligned annotations from different groups

phosphorylation of **p65** at Ser276 prevents its degradation by ubiquitin-mediated proteolysis and promotes cell survival in HeLa cells [24].

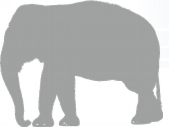
phosphorylation of p65 at Ser276 prevents its degradation by ubiquitin-mediated proteolysis and promotes cell survival in HeLa cells [24].

phosphorylation of p65 at Ser276 prevents its degradation by ubiquitin-mediated proteolysis and promotes cell survival in HeLa cells [24].



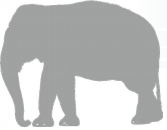
Alignment

- Aligned annotations
 - ✓ <http://www.pubannotation.org/>



Global addressing system

- Persistently preserve the texts of all the articles from PubMed / PMC(OA)
 - ✓ UTF-8
 - ✓ ASCII conversion is provided
- Offset indices are stably maintained



a case of Google Map

← → ↺ <https://www.google.de/maps/search/restaurants/@48.1516079,11.5576419,15z> ☆ font ↻ ⋮

Apps Google DBCLS 네이버 사전 Ontology BLAH BioNLP-ST Virtuoso AIDIC okbqa BioHackathon Other bookmarks

☰ restaurants 🔍 ✕

Rating
★★★★★

Block House München
4.3 ★★★★★ (172)
Steak · Leopoldstraße
Opens at 11:30

Vapiano München 5 Höfe
3.7 ★★★★★ (161)
Italian · Theatinerstraße
Sleek chain for self-serve Italian fare
Opens at 10:00

Hotel Sofitel Munich Bayerpost €392
4.2 ★★★★★
5-star hotel · Bayerstraße
Posh option with elegant rooms & suites

Lemar
4.2 ★★★★★ (39)
Afghani · Viktor-Scheffel-Straße
Opens at 17:00

Restaurant Rila
4.3 ★★★★★ (14)

About pricing ⓘ Showing results 1 - 20 < >

☒ Update results when map moves

Map data ©2016 GeoBasis-DE/BKG (©2009), Google Terms Send feedback 200 m



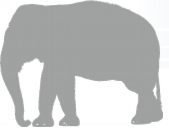
A case of PubAnnotation

- Example of URL
 - ✓ <http://pubannotation.org/docs/sourcedb/PubMed/sourceid/10022882/spans/606-710/annotations/visualize>
- How to get the URL (Example)
 - ✓ <http://pubannotation.org/docs/sourcedb/PubMed/sourceid/10022882>



Global addressing system

- Persistently preserve the texts of all the articles from PubMed / PMC(OA)
 - ✓ UTF-8
 - ✓ ASCII conversion is provided
- Offset indices are stably maintained

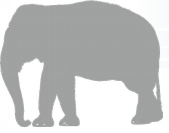


Exchange

Hi Bill, what is the diagnostic test for MERS infection?

Check this [link](#) out.

FYR, I've annotated it using NCIt, OBI, and SNOMEDCT.



Exchange

Hi Bill, what is the diagnostic test for MERS infection?

Check this link out.
FYR, I've annotated it using NCI, OBI, and SNOMEDCT.



TextAE



C17237

C96966

OBI_0000911

C15263

C18020

In June 2012, the U.S. Food and Drug Administration authorized emergency use of the rRT-PCR assay panel as an in vitro diagnostic test for

697932005

MERS-CoV.

Source: <http://pubannotation.org/projects/example/docs/sourcedb/PubMed/sourceid/24153118/spans/1154-1302/annotations.json>



Currently, in PubAnnotation



Make your annotation public, and more useful!

English 日本語 [signup](#) [login](#)

[REPOSITORY](#) [SEARCH](#) [NEWS](#) [GUIDES](#) [ABOUT](#)

[> top](#)

Documents

PMC	9,044	<input type="text" value="keywords"/>	<input type="button" value="Q"/>	<input type="text" value="source ID"/>	<input type="button" value="📄"/>
PubMed	8,424,657	<input type="text" value="keywords"/>	<input type="button" value="Q"/>	<input type="text" value="source ID"/>	<input type="button" value="📄"/>
FirstAuthor	8	<input type="text" value="keywords"/>	<input type="button" value="Q"/>	<input type="text" value="source ID"/>	<input type="button" value="📄"/>

Projects (146)

With most annotations

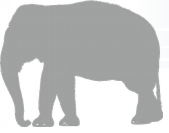
Name		# Ann.	Updated At	Status
PubmedHPO	⚙️	12,437,742	2016-01-31	Uploading
DisGeNET	⚙️	3,117,504	2016-01-28	Beta
NEUROSES	⚙️	2,151,082	2016-02-24	Beta
CRAFT-treebank	👍	844,123	2015-11-19	Beta
bionlp-st-ge-2016-spacy-par...	⚙️	225,680	2016-05-25	Released
spacy-test	⚙️	136,597	2016-05-25	Released
FSU-PRGE	⚙️	59,505	2016-05-17	Released
craft	👍	52,960	2015-10-13	Beta
jnlpba-st-training	👍	51,290	2016-09-09	Released
PennBioIE	👍	23,881	2016-05-17	Released

News

- (09 Mar 2016) Recent access problem
- (25 Feb 2016) System problem on 25/02/2016
- (19 Jan 2016) a new project status, Uploading, added.
- (04 Jan 2016) News service begins.

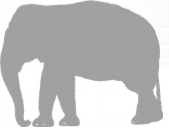
Recently updated

Name		# Ann.	Updated At	Status
jnlpba-st-training	👍	51,290	2016-09-09	Released
jnlpba-st-test	👍	6,005	2016-09-07	Uploading
bionlp-st-seedev-2016-training	👍	109	2016-09-05	Uploading
bionlp-st-bb3-2016-training	👍	1,292	2016-09-05	Released
bionlp-st-cg-2013-training	👍	10,935	2016-08-22	Released
bionlp-st-pc-2013-traing	👍	7,855	2016-08-22	Released
bionlp-st-id-2011-training	👍	5,609	2016-08-22	Released
bionlp-st-epi-2011-training	👍	7,595	2016-08-22	Released
Ab3P-abbreviations	👍	2,342	2016-07-29	Beta
AnEM_full-texts	👍	689	2016-07-27	Uploading



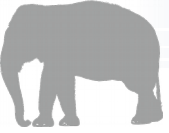
PubAnnotation

- New version to be released soon
 - ✓ Performance improved
 - ✓ Interface improved
 - ✓ BioC conversion to be supported
 - ➔ Thanks to the NCBI team
 - ✓ Bug fixes



TextAE

- To access/edit annotations
 - ✓ <http://textae.pubannotation.org>
 - ✓ Has fully RESTful APIs



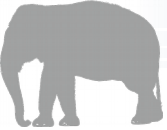
PubDictionaries

- To share dictionary resources
 - ✓ Current version
 - ➔ <http://pubdictionaries.org>
 - ✓ New version
 - ➔ <http://new.pubdictionaries.org>

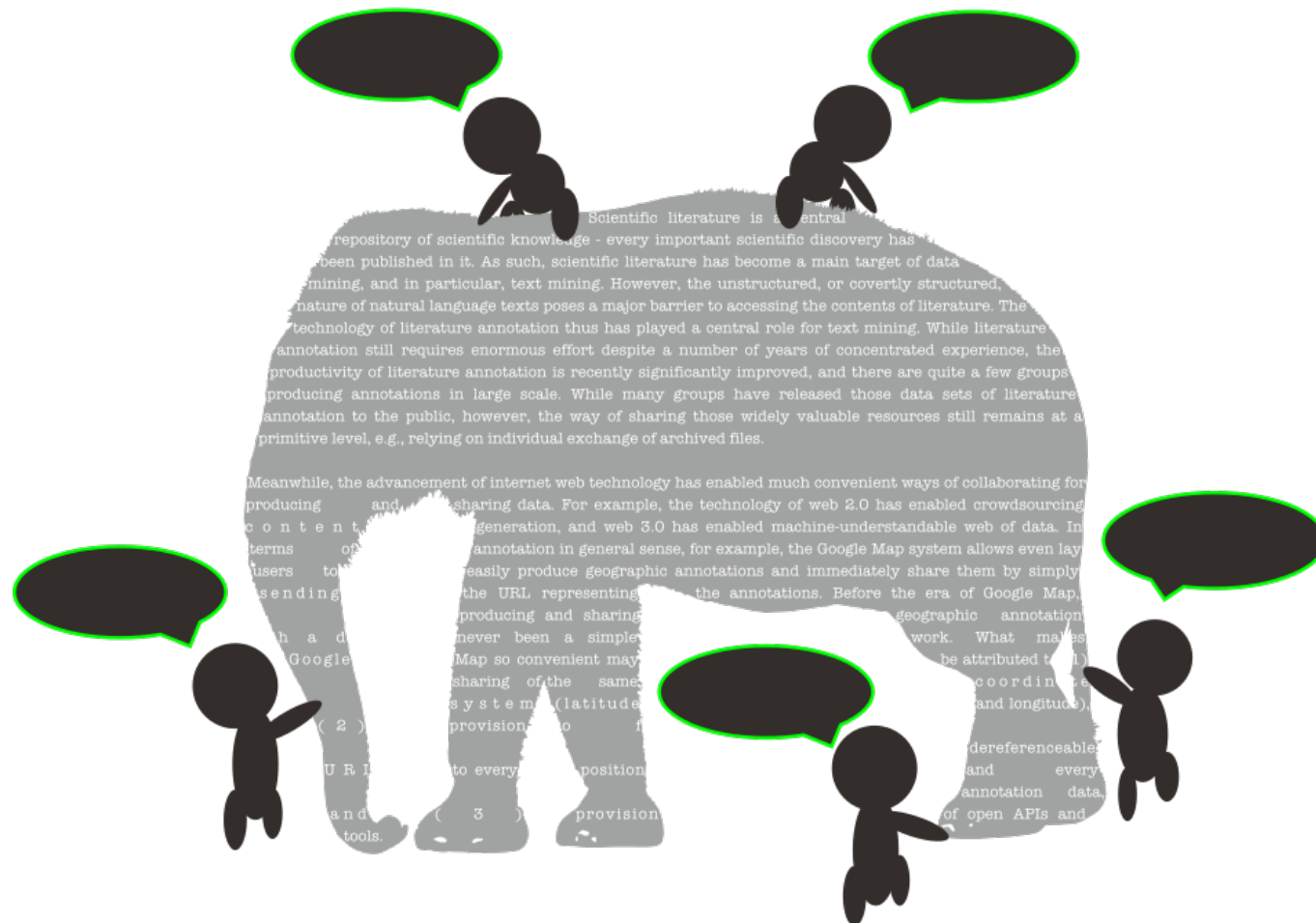
Shared Annotation Targets

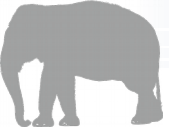
Scientific literature is a central repository of scientific knowledge - every important scientific discovery has been published in it. As such, scientific literature has become a main target of data mining, and linguistic structure of machine-readable structured, unstructured, and semi-structured texts poses a major barrier to accessing the contents of literature. The technology of literature annotation thus has played a central role for text mining. While literature annotation still requires enormous effort despite a number of years of concentrated experience, the productivity of literature annotation is recently significantly improved, and there are quite a few groups producing annotations in large scale. While many groups have released those data sets of literature annotation to the public, however, the way of sharing those widely valuable resources still remains at a primitive level, e.g., relying on individual exchange of archived files.

Meanwhile, the advancement of internet web technology has enabled much convenient ways of collaborating in producing and sharing data. For example, the technology of web 2.0 has enabled crowdsourcing in content generation, and web 3.0 has enabled machine-understandable web of data. In terms of literature annotation in general sense, for example, the Google Map system allows even lay users to easily produce geographic annotations and immediately share them by simply sending the URL representing the annotations. Before the era of Google Map, producing and sharing geographic annotations has never been a simple work. What makes Google Map so convenient may be attributed to the sharing of the same coordinate system (latitude and longitude provision) to all users. A URL to every position and every annotation

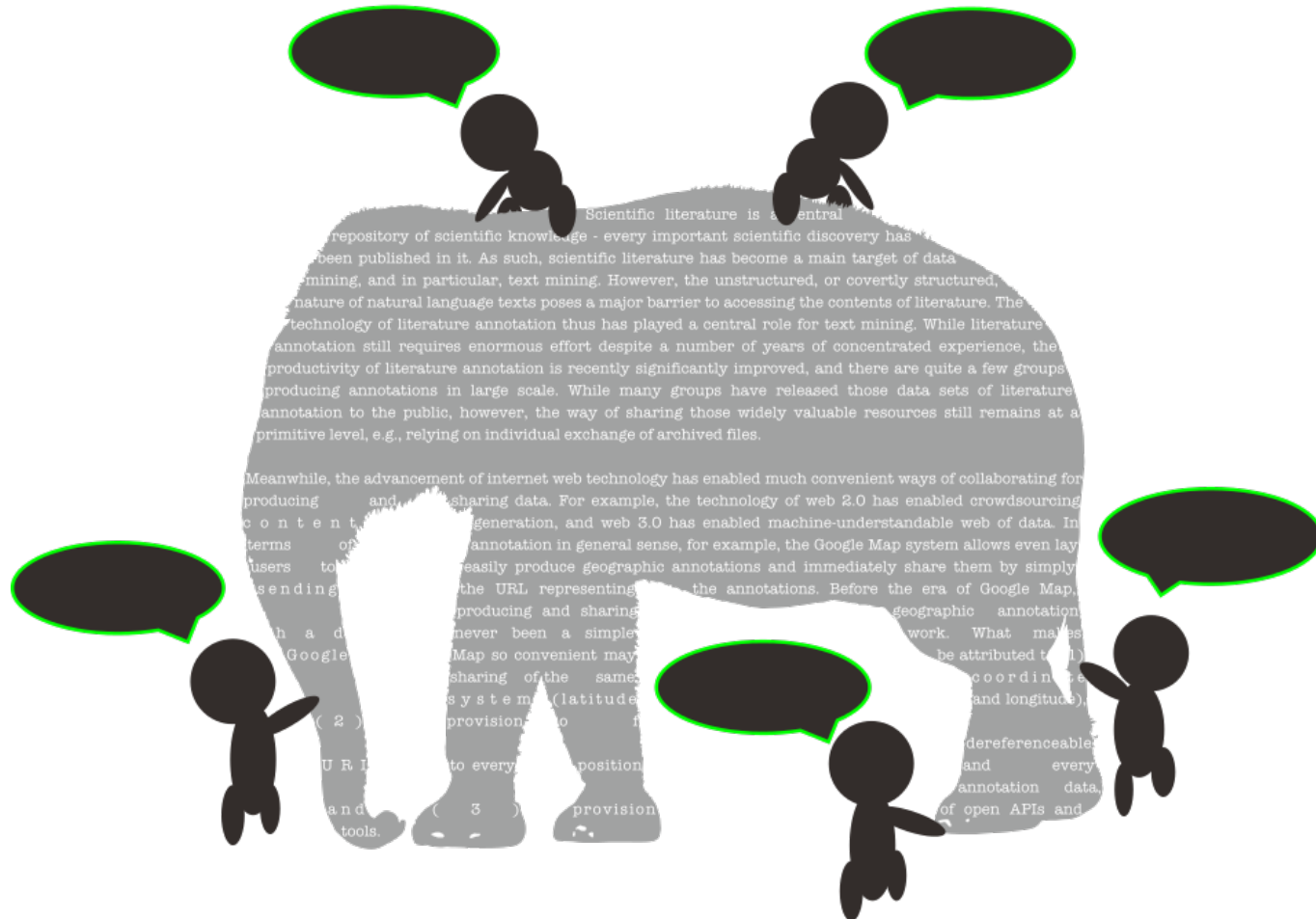


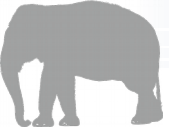
Many literature annotation projects





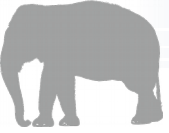
None of them is complete



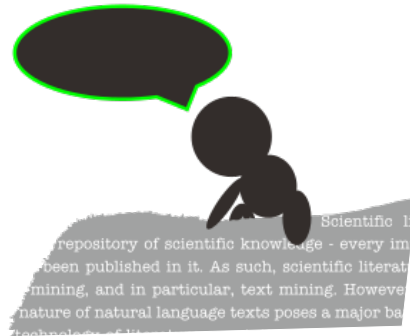


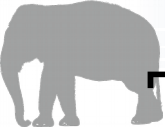
None of them is complete



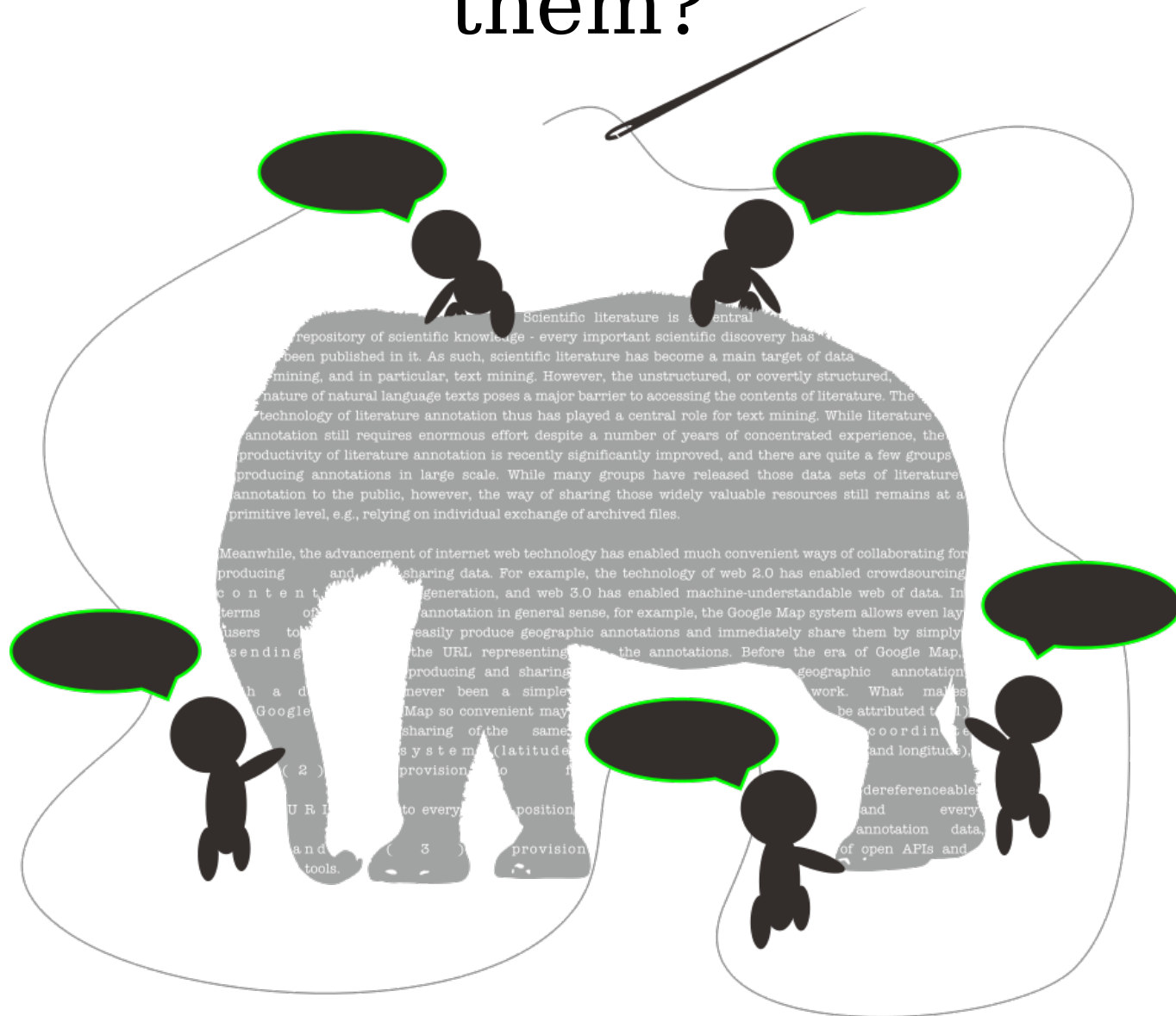


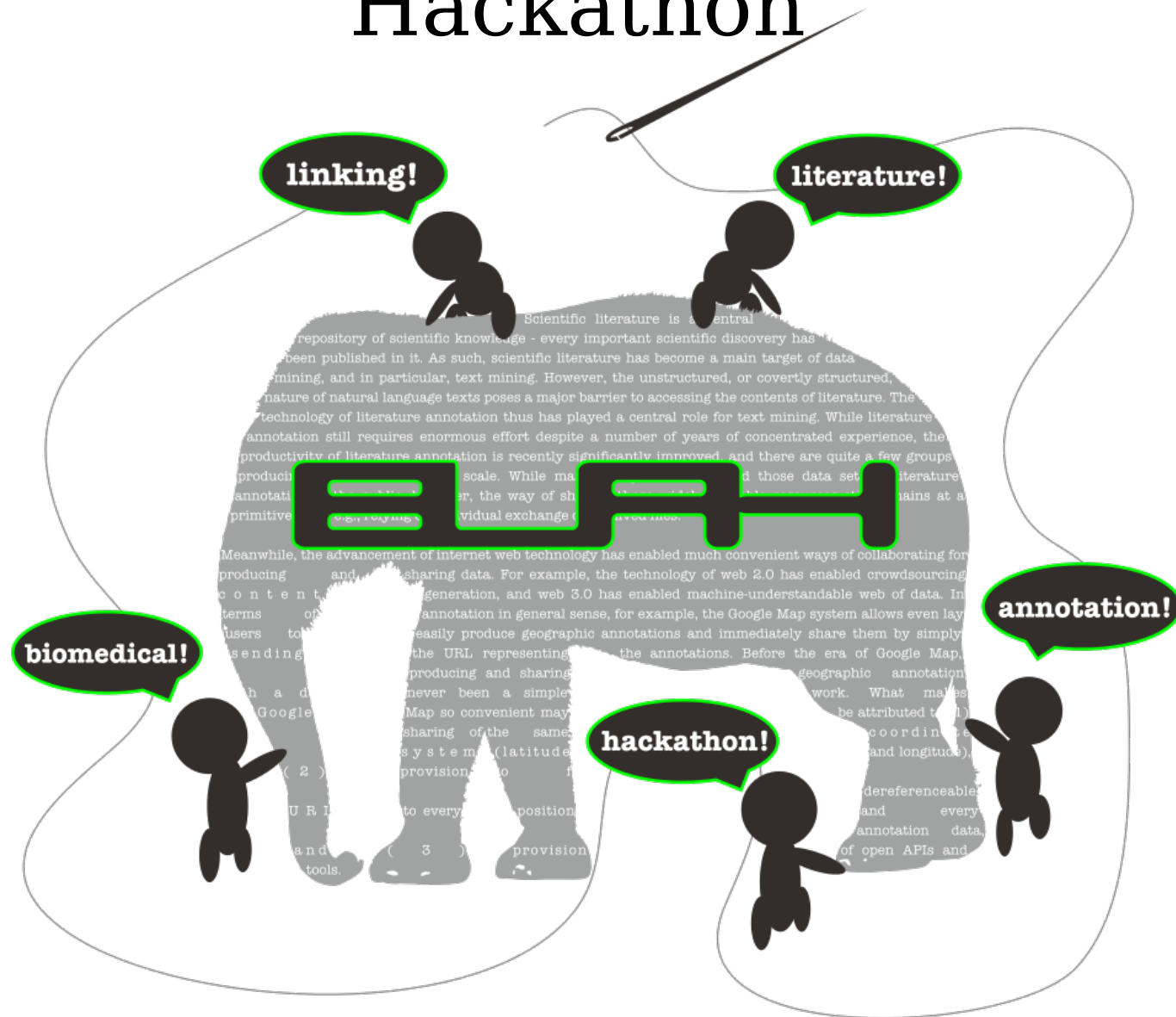
None of them is complete

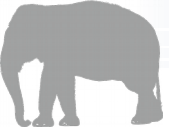




Then, why don't we collect & link them?

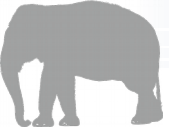






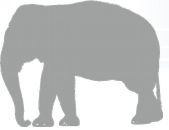
BLAH

- Biomedical Linked Annotation Hackathon
 - ✓ BLAH
 - ➔ Feb. 2015, Kashiwa
 - ✓ BLAH2
 - ➔ Nov. 2015, Mishima / Ito
 - ✓ **BLAHMUC**
 - ➔ **Oct. 2016, Munich**



BLAH

- Biomedical Linked Annotation Hackathon
 - ✓ BLAH
 - ➔ Feb. 2015, Kashiwa
 - ✓ BLAH2
 - ➔ Nov. 2015, Mishima / Ito
 - ✓ BLAHMUC
 - ➔ Oct. 2016, Munich
 - ✓ BLAH3
 - ➔ Jan. 2017, Tokyo



Thank you!
Happy October Blah!