The BioASQ challenge: Resources and Evaluation for Semantic Extraction from the Biomedical Literature

Georgios Paliouras
Acting Director

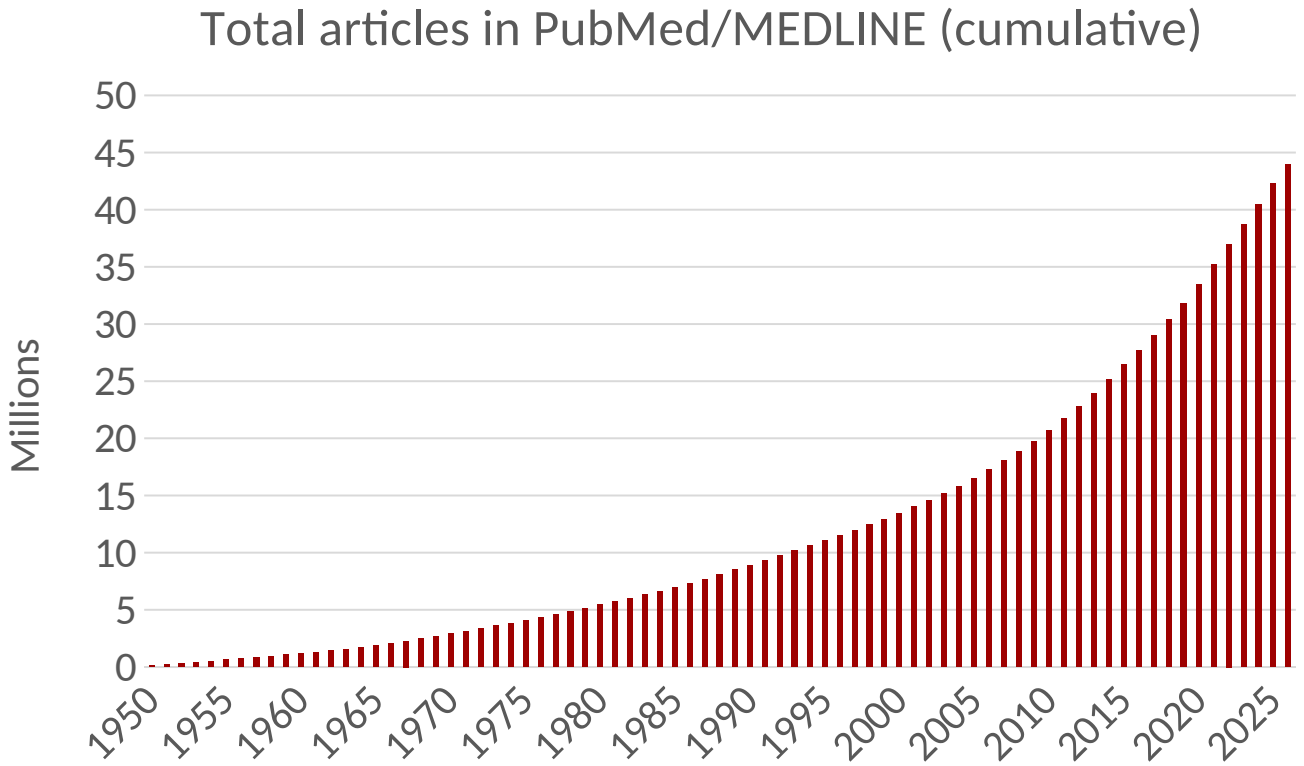
DEMOKRITOS

INSTITUTE OF INFORMATICS & TELECOMMUNICATIONS

OLYMPIA AI DATA INFRASTRUCTURE SUMMIT,
Nov 4-5, 2026 @FORTH, Crete, Greece

The vision

- **More than three** biomedical articles are published every minute!
- Biomedical knowledge should benefit patients and be accessible to biomedical experts
- Push research in information retrieval and question answering through a competition



Sources: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html, http://dan.corlan.net/

What is BioASQ?



- A series of **challenges** on **biomedical semantic indexing** and **question answering** (QA).
- Shared tasks where participants are asked to:
  - **Index** large-scale biomedical resources (i.e. MEDLINE),
  - **Retrieve information** related to natural language questions,
  - **Compose answers** to natural language questions.
- Additional shared tasks with focus on
  - **Developing topics**, such as COVID-19, and
  - **Multilingual, specialty-specific information extraction**, clinical coding and summarization

What is BioASQ?



BioASQ has run fourteen times, in **fourteen years**.

- **14 workshops** in CLEF, ACL, EMNLP, and ECML-PKDD.
- More than **180 teams** from **41 countries**.

BioASQ Tasks Across the Years

Overview of all task editions

Task	Language	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027*
Task A	EN	1	2	3	4	5	6	7	8	9	10					
Task C	EN					1										
MESINESP	ES								1	2						
DisTEMIST	ES										1					
MedProcNER	ES											1				
MultiCardioNER	ES, EN, IT												1	2		
BioNNE(-L/R)	RU, EN												1	2	3	
ELCardioCC	EL													1	2	3
GutBrainIE	EN													1	2	3
Task B	EN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Synergy	EN									1	2	3	4	5	6	7
MultiClinSum	15 languages													1	2	3
PQASD	EN															1

** planned for BioASQ15. Red: semantic indexing and information extraction; blue: question answering and summarization*

The BioASQ Ecosystem

Data sources include both **text** and **structured** info.

- PubMed articles.
- Clinical case reports.
- Greek discharge letters.
- NEREL-BIO dataset.
- ICD-10 codes.

Semantic indexing and extraction

Task A: 16M articles, 30K MeSH labels

MESINESP: 369K articles

DisTEMIST, MedProcNER: 1,000 case reports each

ELCardioCC: 3,000 letters

BioNNE(-L/R): 800 abstracts

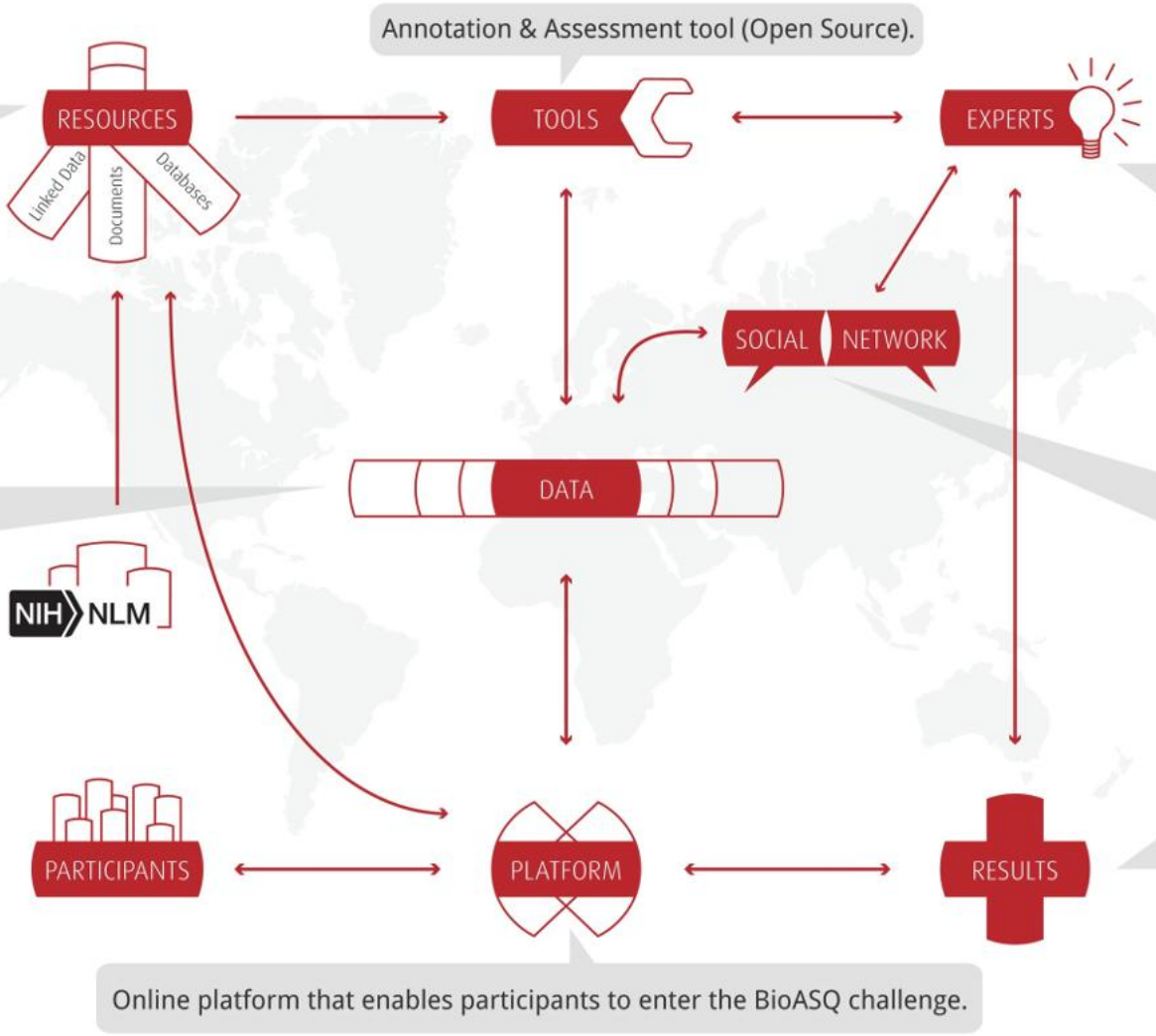
GutBrainIE: 5,000+ abstracts

Question answering and summarization

Task B: 6,009 questions with manual annotations

Synergy: 463 questions with incremental annotations

MultiClinSum: ~26K pairs per language



Biomedical experts, responsible for the formulation of questions, the identification of gold relevant documents, snippets, concepts, triples, the composition of gold "exact" and "ideal" answers, and the manual evaluation of system responses.

A social network allows biomedical experts to follow and comment upon benchmark questions and gold answers formulated by their peers.

Automated & manual assessment of system responses. Oracles available for objective retrospective bench-marking.

BioASQ through the Summit's Lens

Agent Memory Infrastructure

External knowledge as memory of LLM agents

- RAG and agentic pipelines dominate biomedical QA
- Synergy: knowledge accumulates across rounds of expert feedback
- Open question: benchmarking persistent agent memory

Enterprise Ontology

Meaning comes from shared domain vocabularies

- MeSH, ICD-10 and UMLS ground documents in concepts
- Recognition, linking and relations turn text into a knowledge graph
- New PQASD task: AI agents querying structured data

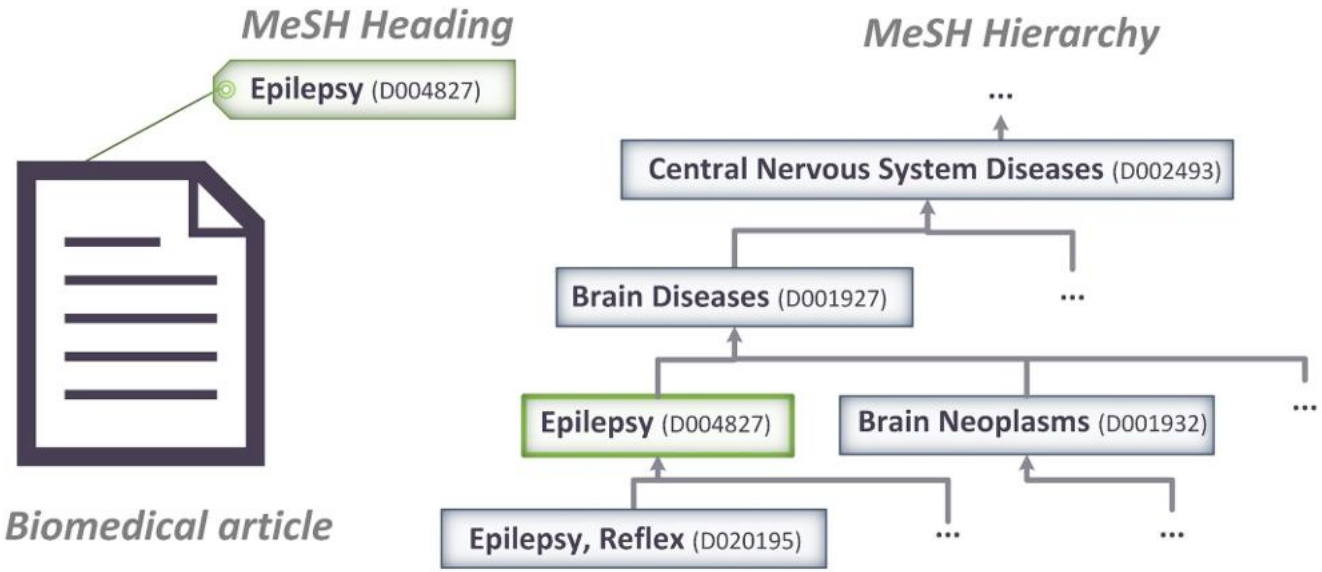
The Future of AI Data Infrastructure

Open benchmarks as a shared AI Data Infrastructure

- Datasets, a shared-task platform and standard evaluation make the progress of AI measurable
- Unstructured literature becomes knowledge-rich, AI-ready resources

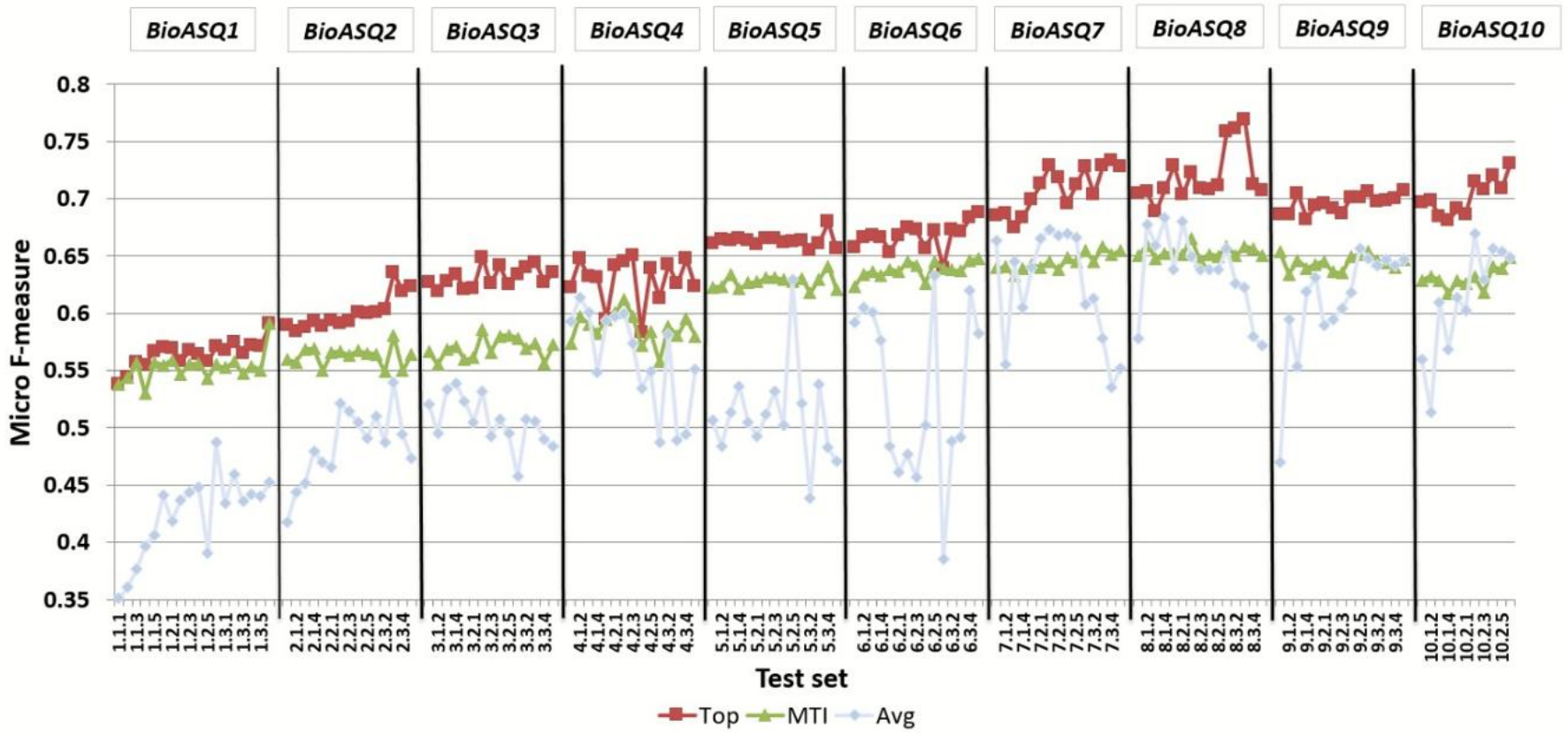
Task A: Biomedical Semantic Indexing

- Extreme Multilabel Classification of MEDLINE articles to Medical Subject Headings (MeSH)
  - **Benchmark dataset**: more than **16M articles** with almost **30K MeSH labels**.
- **Best methods:**
  - **Document representation**: from BOW and tf-idf towards neural embeddings (Doc2Vec) and pre-trained transformers (BioBERT, PubMedBERT)
  - **Concept matching**: from k-NN, SVM and learning-to-rank (MeSHLabeler, DeepMeSH) towards deep networks and attention (AttentionXML, BERTMeSH, neural ranking)



Task A: Biomedical Semantic Indexing

- MTI improved by almost 10%, largely due to BioASQ results. High point: NLM adopted **fully automated indexing** in 2022*.



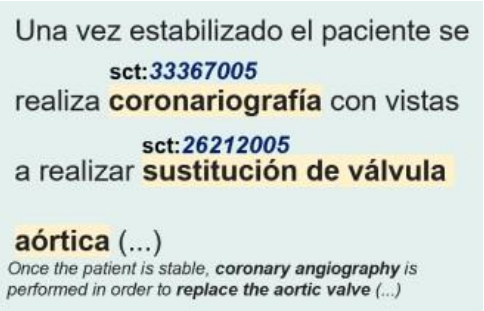
A shared vocabulary (MeSH) is the semantic layer that makes millions of articles consistently findable and reusable

* *https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html*

Semantic Indexing Beyond the English Biomedical Literature

From MESINESP to ELCardioCC: mapping health text to shared vocabularies

- Non-English health text turned into coded, linkable data
  - **MESINESP**: semantic indexing of Spanish literature with DeCS descriptors
  - **DisTEMIST**: NER and SNOMED CT linking for diseases in Spanish clinical cases
  - **MedProcNER**: NER, SNOMED CT linking and document indexing for clinical procedures
  - **MultiCardioNER**: NER for diseases (cardiology adaptation) and drugs in ES, EN and IT
  - **ELCardioCC**: ICD-10 coding of Greek cardiology discharge letters



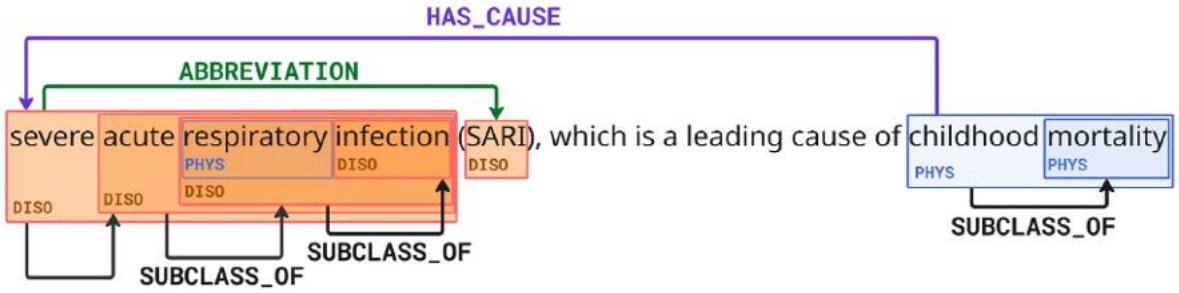
- **Best methods**: multilingual BERT with AttentionXML for indexing, transformers (RoBERTa, SapBERT) for NER and linking, ensembles of Greek BERT and XLM-RoBERTa for ICD-10 coding

- Limited task-specific resources pose additional challenges
  - Need for knowledge transfer across languages, document types, entity types, and medical domains
  - Use of high-quality multilingual shared resources, such as hierarchies and translated vocabularies

Shared vocabularies (DeCS, SNOMED CT, ICD-10) turn clinical text in many languages into interoperable data

Information Extraction: from Entities to Relations

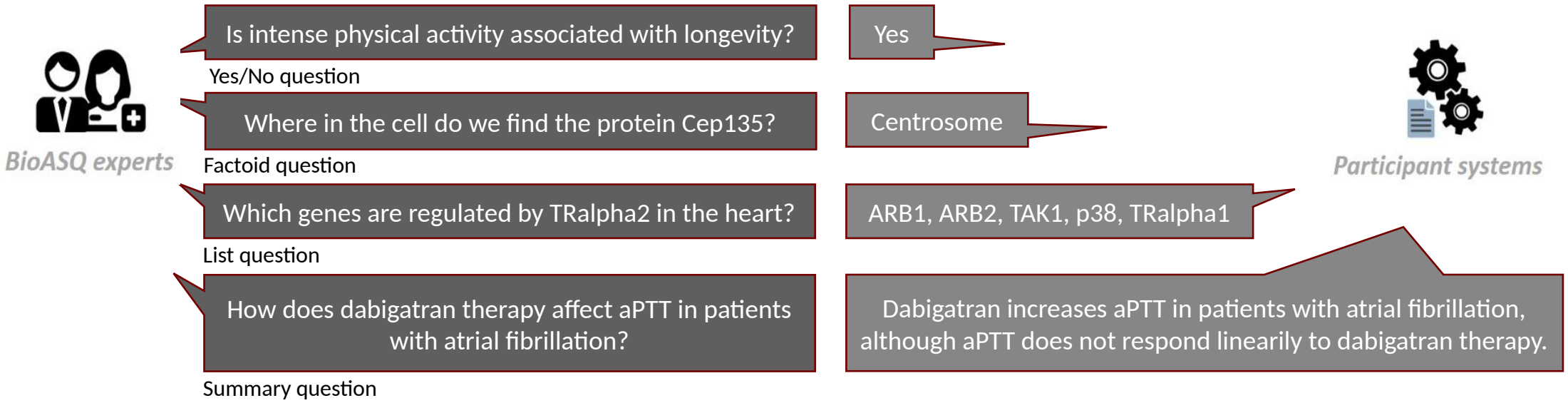
Nested entities in Russian and English, and IE in the domain of the gut-brain axis



- Entities linked to concepts and relations among them convert the literature into structured knowledge
  - **BioNNE**: focus on nested entities in Russian and English biomedical literature
  - **GutBrainIE**: focus on the literature for the gut-brain axis
- **Trend**: from entity taggers, to retrieve-and-rerank linking, to relation classifiers with typed markers and graph models
- **Conclusions**:
  - fine-tuned transformer encoders performed well, while LLM prompting lagged behind
  - performance drops from entities to concepts to relations
  - nested structure and rare types remain a challenge

linking extracted entities and relations to shared vocabularies (e.g. UMLS) turns text into a knowledge graph

Task B: Biomedical Question Answering



- **Information Retrieval**: find relevant articles and snippets (Phase A)
- **Question Answering**: produce "exact" (short) and "ideal" answers (summaries) with access to ground-truth relevant material (Phase B) or without (Phase A+)
- **Benchmark dataset**: **6,009** biomedical **questions**, curated **manually**

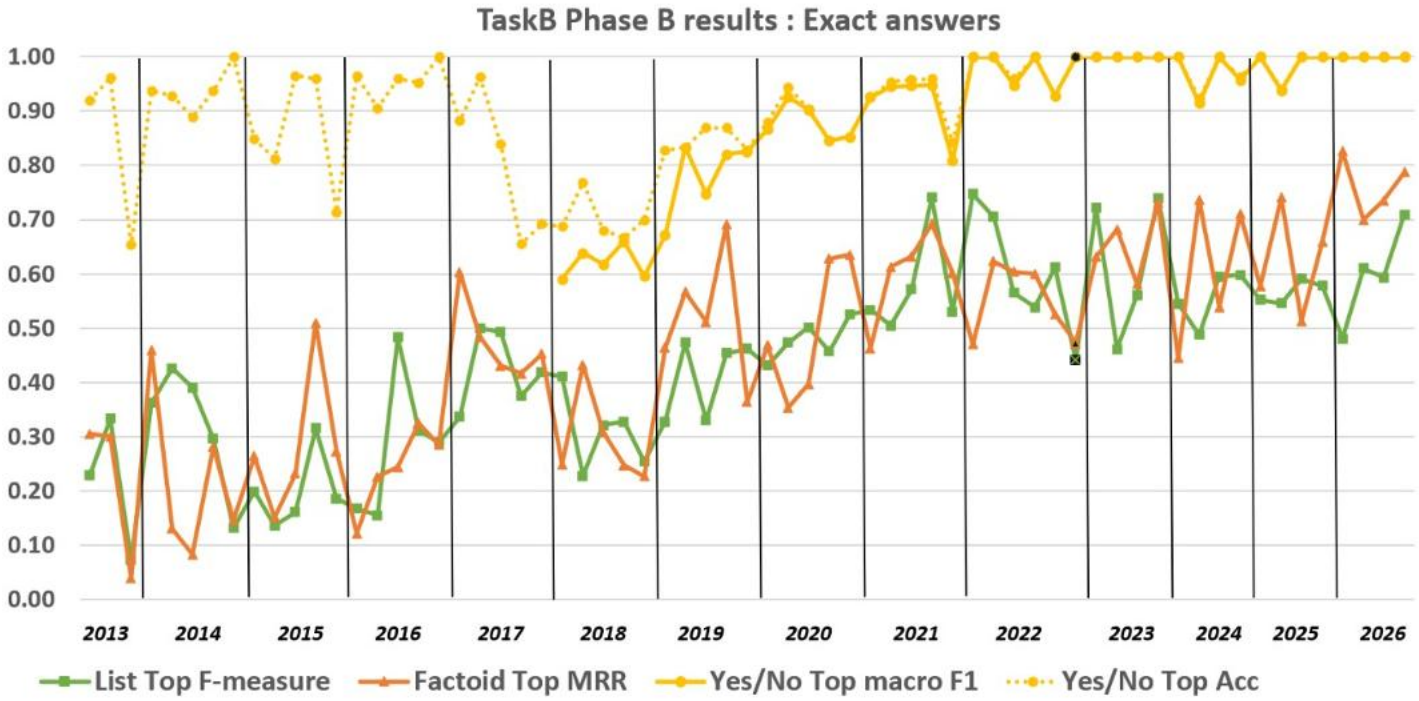
Retrieved external knowledge acts as the memory of LLM agents

Open datasets and a shared platform make progress in this task measurable

Task B: Participant Methods

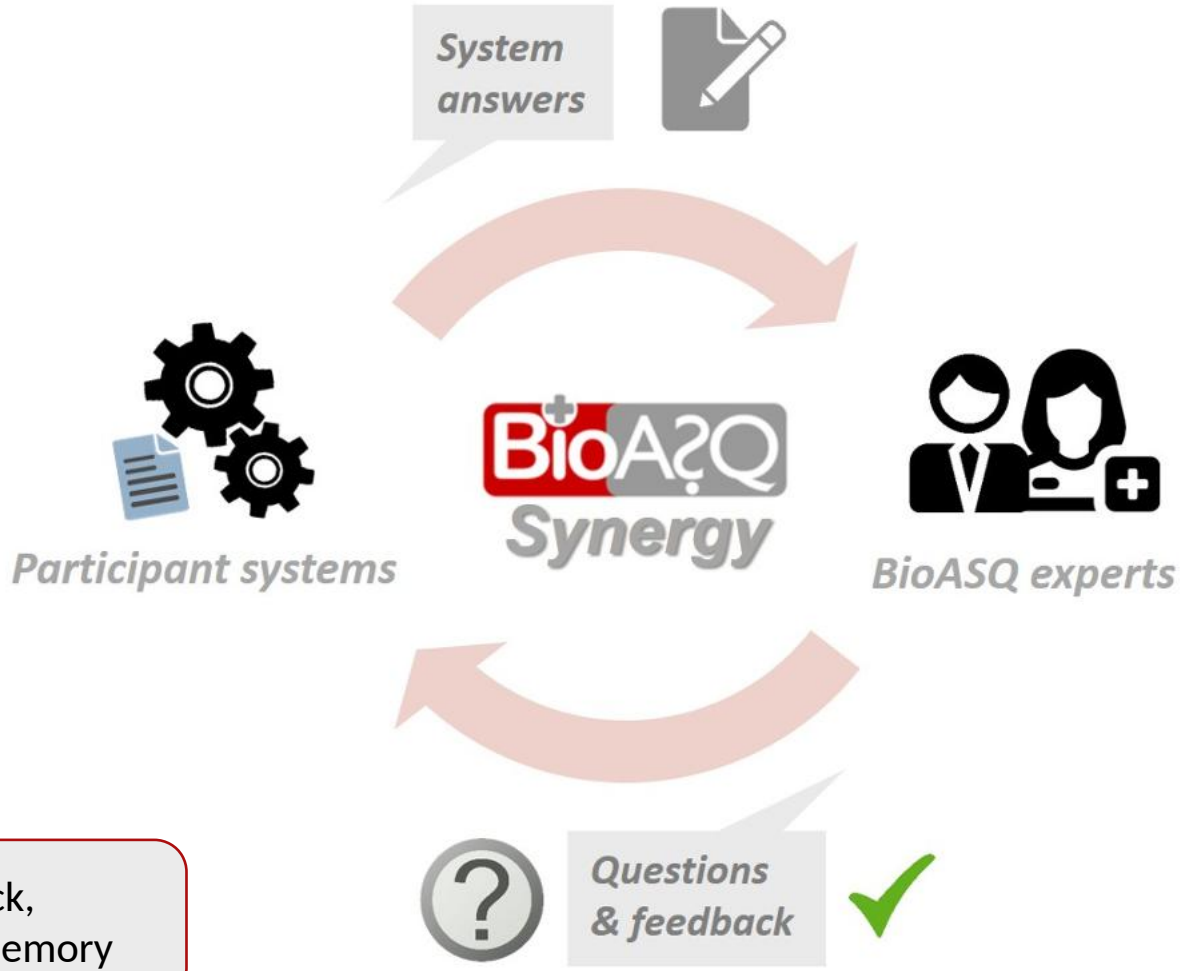
Evolution across fourteen editions

- **Best methods**: trend from structured pipelines with BOW and ontology-based NER, to neural embeddings, LLM-based RAG, and agentic pipelines
  - **Hybrid retrieval**: BM25 plus dense models (BGE-M3, MedCPT, SPLADE) fused with RRF
  - **Neural re-ranking**: cross-encoders, some fine-tuned on BioASQ data
  - **Query expansion** with MeSH and entities, combining models with external knowledge
  - **RAG generation**: GPT, Gemini, Llama, Qwen, and MedGemma with few-shot prompting
  - **Agentic and multi-agent**: agent quorum, ReAct, LLM-as-judge



Task Synergy: Answering Open Research Questions

- A dialogue between systems and supporting **incremental understanding**
  - Open questions that may lack a definite answer
  - Developing health topics such as COVID-19
  - Evolving document collection
- **Trend:** Task B pipelines reused, with iterative retrieval and diversity-aware evidence selection



Knowledge accumulates across rounds of expert feedback, raising the question of benchmarking persistent agent memory

QA Beyond the English Biomedical Literature

Multilingual clinical case summarization, and pharmaceutical QA over structured data

- **MultiClinSum**: multilingual summaries of long clinical case reports support clinicians and data reuse
  - Summarization in 15 languages: EN, ES, FR, PT, IT, RU, CA, NO, DA, RO, DE, EL, NL, CS, SV
    - Multi-agent LLM-as-judge Evaluation: summary consistency across languages is challenging
  - **New in 2027**: entity-aware clinical summarization

- **New in 2027** **PQASD**: pharmaceutical QA grounded over structured and reliable data supports drug development
  - AI agents translate questions into GraphQL queries for the Open Targets API, rather than recalling facts

Trusted ontologies and knowledge graphs ground LLM summaries
Agents should answer from live structured data querying a shared semantic schema

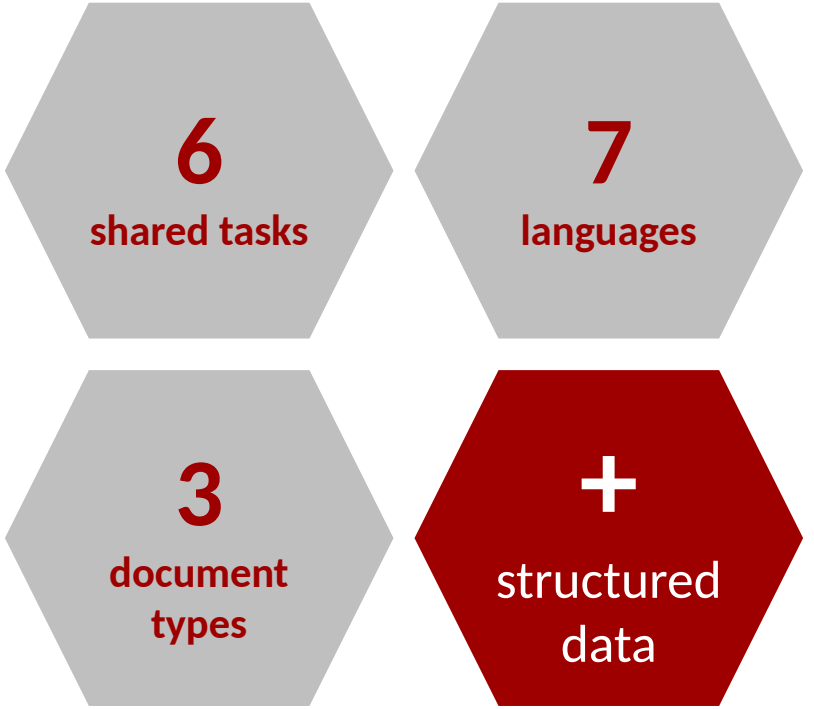
BioASQ Plans for 2027…

New editions of existing tasks on:

- Biomedical semantic QA (**Task B**)
- QA on developing topics (**Synergy**)
- Clinical coding of Greek discharge letters (**ELCardioCC**)
- Gut–brain information extraction (**GutBrainIE**)

New and modified tasks:

- Entity-aware clinical summarization (**MultiClinSum-3**)
- **New task:** pharmaceutical QA over structured data (**PQASD**)



BioASQ15 workshop at CLEF 2027, Bucharest, Romania, 14–17 September 2027

Looking Back...

What fourteen years of BioASQ taught us

- **Stability** is key to trust by community: 14 editions, 180+ teams from 41 countries
- **Adaptation** should be anticipated: tasks were added and retired as technology and needs changed
- **Controlled innovation** keeps the interest and expands the community
- **Real-world tasks can be hard** but can also **shape the real world**: NLM moved to fully automated MEDLINE indexing
- **Live benchmarks pay a price**: evolving data sources and vocabularies complicate long-running tasks
- **Experts and participants steer the tasks**: cross-disciplinary challenges should be exposed to all related communities.
- **Sustainability** requires automation, trust, impact and planning

Shared infrastructure, trusted vocabularies and continuous expert feedback make domain AI measurable

...and Into the Future

BioASQ15 at CLEF 2027, Bucharest, 14-17 September 2027

- **Semantic indexing and extraction**:
  - Fine-grained indexing (concept-level)
  - Multilinguality and low-resource languages
  - Specialized biomedical domains
- **Question answering**:
  - Agentic and retrieval-augmented systems
  - Use of structured resources for grounding answers
  - Entity-aware summarization
- **Evaluation**:
  - LLM-as-judge validated by experts, evidence and attribution
  - Benchmarks for persistent agent memory
  - Community-driven data curation
- **Open systems**: promote openly available systems and reproducibility

Lessons for AI Data Infrastructure

What BioASQ shows for each Summit theme, and the opportunity ahead

	What BioASQ shows	Opportunity
Agent Memory Infrastructure	• RAG and agentic pipelines dominate QA • Synergy is a continuous expert–system dialogue over an evolving PubMed	Need for benchmarks on persistent, evolving agent memory
Enterprise Ontology	• Concept vocabularies (MeSH, ICD-10, UMLS) are the grounding layer • Errors accumulate from entities to concepts to relations • LLM prompting lags behind supervised models	Related Information Extraction tasks based on enterprise ontologies
The Future of AI Data Infrastructure	• Open, evolving, benchmarks for challenging real-world tasks • Evaluation mixes automatic metrics, expert assessment, and LLM-as-judge	Shared benchmarking infrastructure considering data drift and community involvement

Grateful to our Sponsors…

Gold Sponsor

Ovid®

Past Sponsors

Google

ATYPON

ELSEVIER

NIH NLM

VISEO

www.bioasq.org | @BioASQ

Thank you!

Georgios Paliouras

DEMOKRITOS

INSTITUTE OF INFORMATICS & TELECOMMUNICATIONS