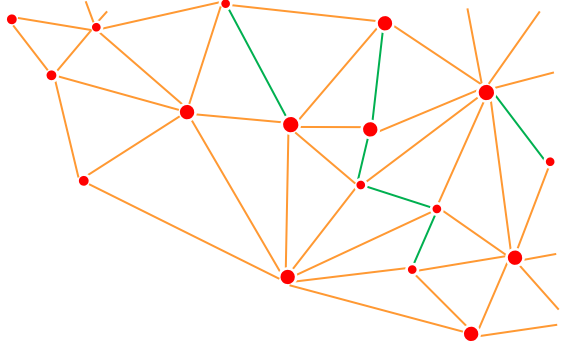
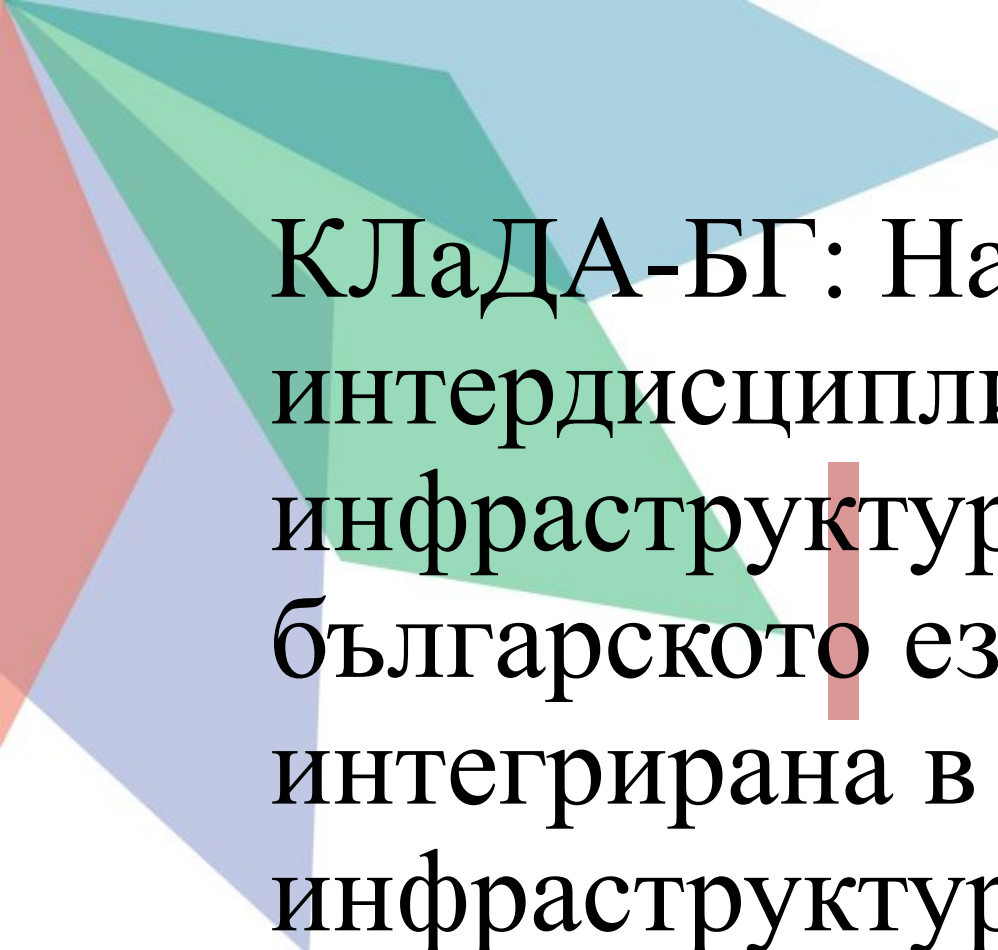


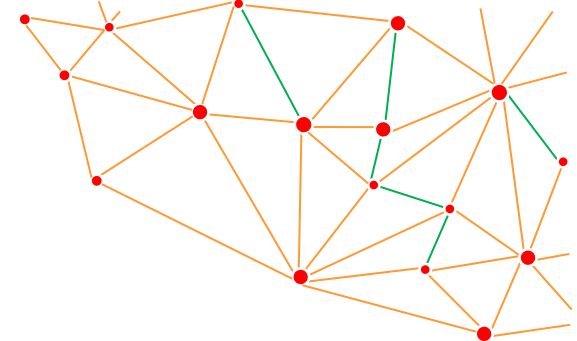


# КЛаДА-БГ: Национална интердисциплинарна изследователска Е- инфраструктура за ресурси и технологии за българското езиково и културно наследство, интегрирана в рамките на европейските инфраструктури CLARIN и DARIAH

Кирил Симов, координатор

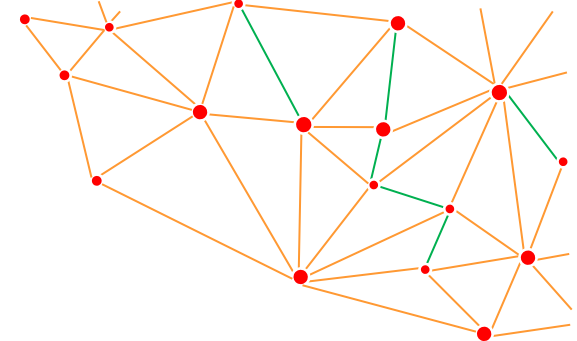
Институт по информационни и комуникационни технологии, БАН

# Мисия на КЛаДА-БГ



- Създаване на виртуална инфраструктура за езикови технологии, културно и историческо наследство
- Езикови технологии за крайния потребител
- Семантични технологии за моделиране на артефакти
- Отворени свързани данни (Linked Open Data) за езикови технологии, културно и историческо наследство
- 3D технологии за сканиране, моделиране, визуализация на културни и исторически артефакти, мултимодални данни
- **Основна цел: подкрепа на изследователската дейност в областта на социалните науки и хуманитаристиката (CH&X)**
- **Прилагане и адаптация на създадените ресурси и технологии към нови области, приложения и задачи**

# Изследователски инфраструктури



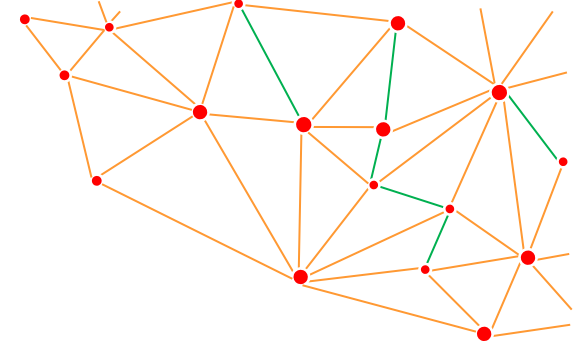
Изследователските инфраструктури са сложни системи,  
включващи:

- **знания,**
- **данни,**
- **хора и**
- **услуги,**

обединяващи различни ресурси за нуждите на **широк кръг от потребители**. Тези ресурси трябва да са **използваеми и достъпни** за достатъчно **дълъг период**, за да **подкрепят научните изследвания** (индивидуални или съвместни) и да **споделят резултатите от тези изследвания** (**FAIR** принципи)

*PARTHENOS Project (H2020 project of DARIAH and CLARIN)*

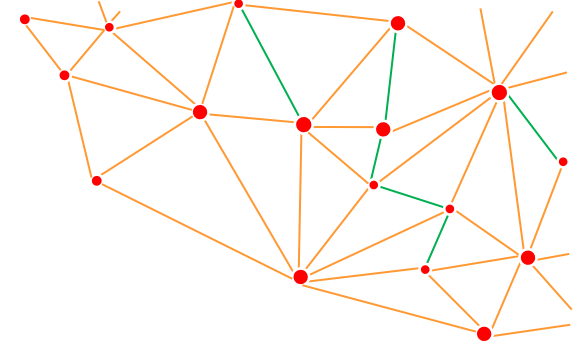
# Данни и знание



- **Данни** – документи в различни формати и медии: текст, хипертекст, аудио, видео, изображения
- **Знание** – интерпретация на данните с оглед на създаване, съдържание и взаимосвързаност
- В КЛаДА-БГ приемаме, че основното количество данни са неструктурирани и не-хомогенни текстови данни: писма, протоколи, записки и други

„Language Technology is a core Big Data technology. Growth in the volume and variety of data is mostly due to the accumulation of unstructured text data; in fact, **up to 80% of all data is unstructured text data.**“ <http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf>  
(p. 110)

# Изследователски процес в CN&X



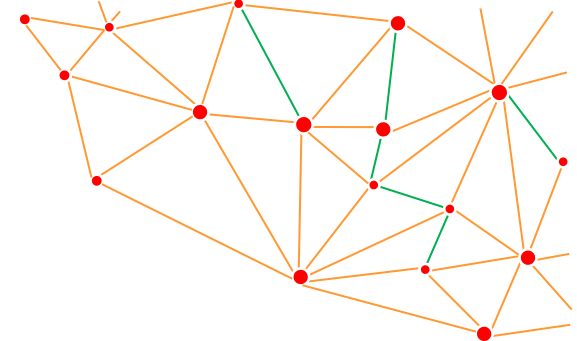
- Огромно многообразие от изследователски обекти
- Тези обекти съдържат информация за едни и същи същности, но не са свързани
- Изследователският процес в CN&X изисква поставянето на тази информация в общ контекст:

Процесът на интеграция на данните *контекстуализация*

- Точността, автентичността, количеството на данните и знанието достъпни по този начин са ключови за използваемостта на инфраструктурата

**Изследователят** трябва да има достъп до цялата налична информация и знание на **едно място**, с множество възможности за наблюдение на тези данни и механизми за извършване на изследователските дейности

# Външно финансирани проекти



## **ParlaMint** *(ИИКТ-БАН)*

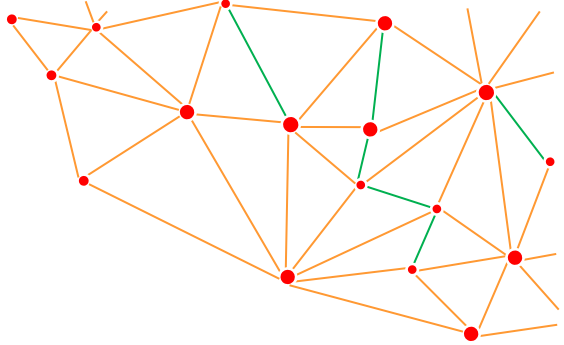
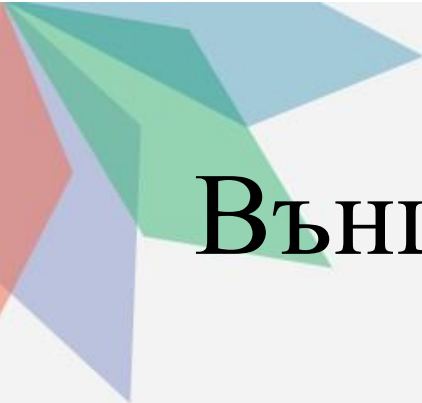
Проект, финансиран от CLARIN-ERIC, който има за цел изработване на многоезиков корпус с парламентарни дебати на много европейски езици. Корпусите са представени в обща структура и съпоставими метаданни. (<https://www.clarin.eu/parlamint>)

## **ParlaCap** *(ИИКТ-БАН)*

Проект, финансиран от Европейския съюз, за разширяване на корпусите, създадени в рамките на ParlaMint. OSCARS Project.  
(<https://oscars-project.eu/projects/parlacap-comparing-agenda-settings-across-parliaments-parlamint-dataset>)

## **PressMint** *(НБИБ-Пловдив, ИИКТ-БАН)*

Проект, финансиран от CLARIN-ERIC, за създаване на корпус от исторически публицистични материали от 19 в. до средата на 20 в. Корпусите ще бъде анотиран с наименования същности. Корпусите ще използват единно представяне и съпоставими метаданни. (<https://www.clarin.eu/pressmint>)



# Външно финансирани проекти

## **„Расте, но не старее: семантични и 3D технологии за историята на улиците и сградите в София“** *(ИБЦТ-БАН, ИИКТ-БАН, РИМ-София)*

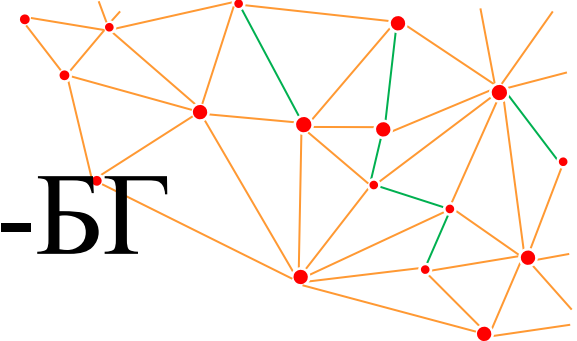
Проектът се фокусира върху проучването и иновативното представяне на историята на град София в периода до средата на XX век. Дейностите предвиждат активно взаимодействие между хуманитарните изследвания, историята на града и съвременните цифрови технологии като средство за създаване и разпространение на ново научно знание, ангажиране с градската среда и адекватно на днешните нужди социализиране на културното наследство. *(финансиран от ФНИ с договор номер КП-06-Н90/2 от 10.12.2024 - <https://historicalsofia.eu/bg>)*

## **Българската онлайн енциклопедия. Свързани и достъпни енциклопедични знания** *(НЦБЕ-БАН, ИИКТ-БАН)*

Проектът разработва иновативна онлайн академична енциклопедична платформа, чрез която да бъде дигитализиран и предоставен за обществено ползване богатият ресурс от научни енциклопедични статии на Научноинформационен център „Българска енциклопедия“. *(BG-RRP-2.017 „Финансиране на научноизследователски проекти в областта на зелените и цифровите технологии - 2“)*



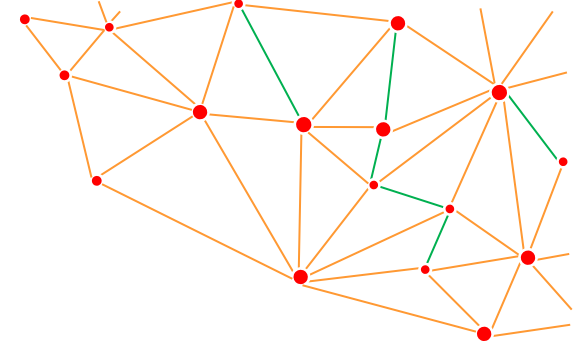
# Ден на отворените врати на КЛаДА-БГ



- Разглеждаме текущото състояние на КЛаДА-БГ като част от **основния период на изграждане на КЛаДА-БГ**, който включва имплементиране на ключови услуги ресурси, идентифициране на техни разширения, работа с потребители и др. Някои от тях ще представим днес
- В работата в рамките на КЛаДА-БГ участват повече от 100 души
- Предстои:
  - Да се имплементират пълни версии на услугите и ресурсите
  - Да успеем да подкрепим изследователската работа в областта на CN&X



# Програма на събитието



09:00 Откриване

Сесия за технологии и ресурси на КЛаДА-БГ

09:15 *Представяне на КЛаДА-БГ* (Кирил Симов, ИИКТ-БАН)

09:30 *Езикови ресурси и технологии, големи езикови модели за обработка на данни на български език* (Петя Осенова, ИИКТ-БАН и СУ)

09:50 *Мрежа от знания, услуги и големи езикови модели за извличане на знание* (Кирил Симов, ИИКТ-БАН)

10:10 *3D технологии, модели и проекти* (Иван Георгиев, ИИКТ-БАН)

10:30 *Кафе пауза*

11:00 *CultIS – софтуерна платформа за интелигентно цифрово управление и представяне на големи масиви от данни и знания от хуманитаристиката и социалните науки. Приложения.* (Десислава Панева-Маринова, ИМИ-БАН)

11:20 *От скрипторий до сървър: дигитални инфраструктури за изследване на културното наследство* (Цветан Василев и Кристиян Симеонов, СУ)

11:40 *Съвременни решения за представяне на културно-историческото наследство: примерът с ръкописното и книжовно наследство на Неофит Бозвели* (Юра Константинова, ИБТЦ-БАН)

12:00 *Уеббазираната система LABMETA – параметри, функционалности* (Жанета Андреева, ШУ)

# Програма на събитието: постери

- **Стефка Александрова.** *PoliteLab* – уеббазирана система за създаване на многоезичен онлайн речник на речевия етикет (разработвана по ННП „Развитие и утвърждаване на българистиката в чужбина“, инспирирана от КлаДа-БГ.)
- **Ани Ангелова.** *PhrasoLAB-BG* – уеббазирана платформа за изучаване на български и немски фразеологични единици
- **Жанета Андреева.** *LABMETA* – уеббазирана система за изследване на когнитивни метафори в българските политически речи
- **Цветан Василев.** *Кирило-методиевската традиция в CLaDA-BG: извори, ръкописи, артефакти.*
- **Валентин Витанов.** *Истории зад фасадите: Жилищна сграда на Захари Алкалай*
- **Елена Водинчар и Галин Георгиев.** *Онлайн библиотека и виртуална карта. Българите в Бесарабия*
- **Теодор Вълчев и Николай Паев.** *Визуализация на анотирани документи с големи езикови модели*
- **Марио Филипов.** *Истории зад фасадите: Домът на д-р Костов*
- **Димитър Минев и Иван Крачанов.** *Дигитална библиотека на НБИБ-Пловдив*
- **Ивайло Начев.** *Расте, но не старее: семантични и 3D технологии за историята на улиците и сградите в София*
- **Десислава Панева-Маринова.** *CultIS* – софтуерна платформа за интелигентно цифрово управление и представяне на големи масиви от данни и знания от хуманитаристиката и социалните науки. Приложения
- **Николай Петков, Миряна Райковска, Миглена Райковска.** *Дигитална визия за древни артефакти: Изследване на керамични съдове за ароматно масло от древна Тракия*
- **Николай Петков, Миряна Райковска, Миглена Райковска.** *Триизмерна виртуална реконструкция на късно елинистичен кернос от Водици, Североизточна България*
- **Милен Петров.** *Исторически хамелеони*
- **Велка Попова, Димитър Попов.** *Български корпус в интерактивната платформа CHILDES*
- **Велка Попова, Димитър Попов.** *Lablass* – уеббазирана система за проучване на словесните асоциации
- **Кристиян Симеонов.** *Осветляване на неизвестното минало на България: Книга с анализите на Филипополската митрополия*
- **Карина Симеонова.** *Истории зад фасадите: Хотел Прентан*
- **Кирил Симов, Петя Осенова, Стефан Маринов, Николай Паев, Кристина Димитрова, Цветелина Александрова.** *Българският корпус със събития: модели за извличане на знания от текст*
- **Кирил Симов, Николай Паев, Петя Осенова.** *Големи езикови модели за преобразуване на предишни варианти на българския правопис към съвременната правописна норма*
- **Яна Янчева и Владимир Кромбърг.** *Спомени от нашето детство*