appen

Data Collection

We have over 20 years' experience completing more than 1,000 bespoke data collection projects across 60+ countries for our customers.

At Appen, we provide data collection services to improve machine learning at scale. Our global footprint allows our clients to quickly capture large volumes of high-quality, customized data, including image, video, speech, audio, and text for their specific AI program needs.

We provide data collection as a standalone service as well as a component of your larger linguistic, transcription, image or audio AI/ML project design. Text, audio, image and video data we collect can be annotated according to your specific guidelines or to our standard conventions.

Fast-track your project and your investment by licensing our off-the-shelf data sets https://appen.com/solutions/data-collection/

Data exists in many places. We are there to find and collect it.

- **Crowdsourced:** We have access to over 1 million skilled contractors, across 130+ countries and in 180+ languages giving you the diversity and scalability you need for high-quality training data. Typically, our crowd uses our proprietary, multi-device mobile app to record audio, image and video data in their home or in public environments as required.
- **Precision Collection:** We offer multi-country, fully supervised data collection sessions using specialized equipment, organized at a central location (e.g. recording studio or a rented home environment).
- **Mass Media:** We collect publicly available online and media data in accordance with the Fair Use clause of your country's copyright laws.
- **Off The Shelf:** We offer licensable speech recognition databases and text corpora, including fully transcribed speech datasets, pronunciation lexicons, POS-tagged lexicons and more, to quickly expand your voice recognition products.
- **Custom Crowd:** Need a specialized crowd? We can use our global recruitment network to find the specialists and domain knowledge you need to help collect your specific data requirements.

Data collection in action with our global crowd of 1 million+ contributors

Analyze
Analyze customer requirements

Design
Design appropriate data collection workflow & quality assurance

Engage
Engage our crowd in the target locales

Collect
Our crowd collects data using an Appen app on their smartphone

Upload
Data is uploaded to Appen servers

Annotate
Data is loaded into the Appen Data Annotation Platform for quality assurance and annotation

Package
Data & metadata packaged according to client requirements & delivered over SFTP

Use Cases for Data Collection

Speech/Audio
For training voice-prompted virtual assistants, voice activated search functions, transcription services, voice-to-text capabilities and more.

Image
For training object recognition, landmark identification, optical character recognition (OCR), biometric identification models, and more.

Video
For training hand gesture and action recognition, biometric identification software, Augmented Reality/ Virtual Reality systems and more.

Why Appen?

Whether you are building highly-specialized applications or deploying international AI-enabled products, our global, enterprise-grade deployment and data collection abilities give you confidence that the volume, quality and velocity of your data requirements are met. At Appen we can partner with you and deliver:

An end-to-end managed service: We cover collection design, large-scale field operations, data quality assurance, and annotation with over 20 years' of deep expertise.

High-quality training data: We have developed methodologies for data quality assurance that take place both during collection and post-collection using our specialized, proprietary data collection tools and quality assurance platforms.

Global coverage of languages and countries: Our crowd has over 1 million skilled contractors in over 130 countries who speak over 180 different languages and dialects. We can ensure your training data meets the demographic needs of your project.

Ethical and secure collection of data: Enjoy peace of mind knowing your data has been collected ethically and in compliance with GDPR and other data security standards. Personally Identifiable Information is managed in accordance with government regulations and all data is stored in secured servers and delivered over SFTP. Our crowd is fairly compensated for the data they provide in accordance with our Fair Pay policy.

Efficient management of metadata: In addition to our bespoke data collection for your specific projects, we also ensure speech data is accurately tagged with the right gender, age, dialect and other demographic details.

Complete data lifecycle services

We also have a variety of other services which can complement and enhance your data collection needs including:

Annotation and linguist support
To help ensure you are collecting the correct type of data

Video transcription services
To accurately accommodate different accents, specific terminology, entity specific closed captioning, colloquialisms and other language structural and phonetic nuances

Text generation services
To generate scenario-based responses or conversations amongst native speakers with optional subsequent semantic annotation to create a text corpus for chatbot training or natural language processing.